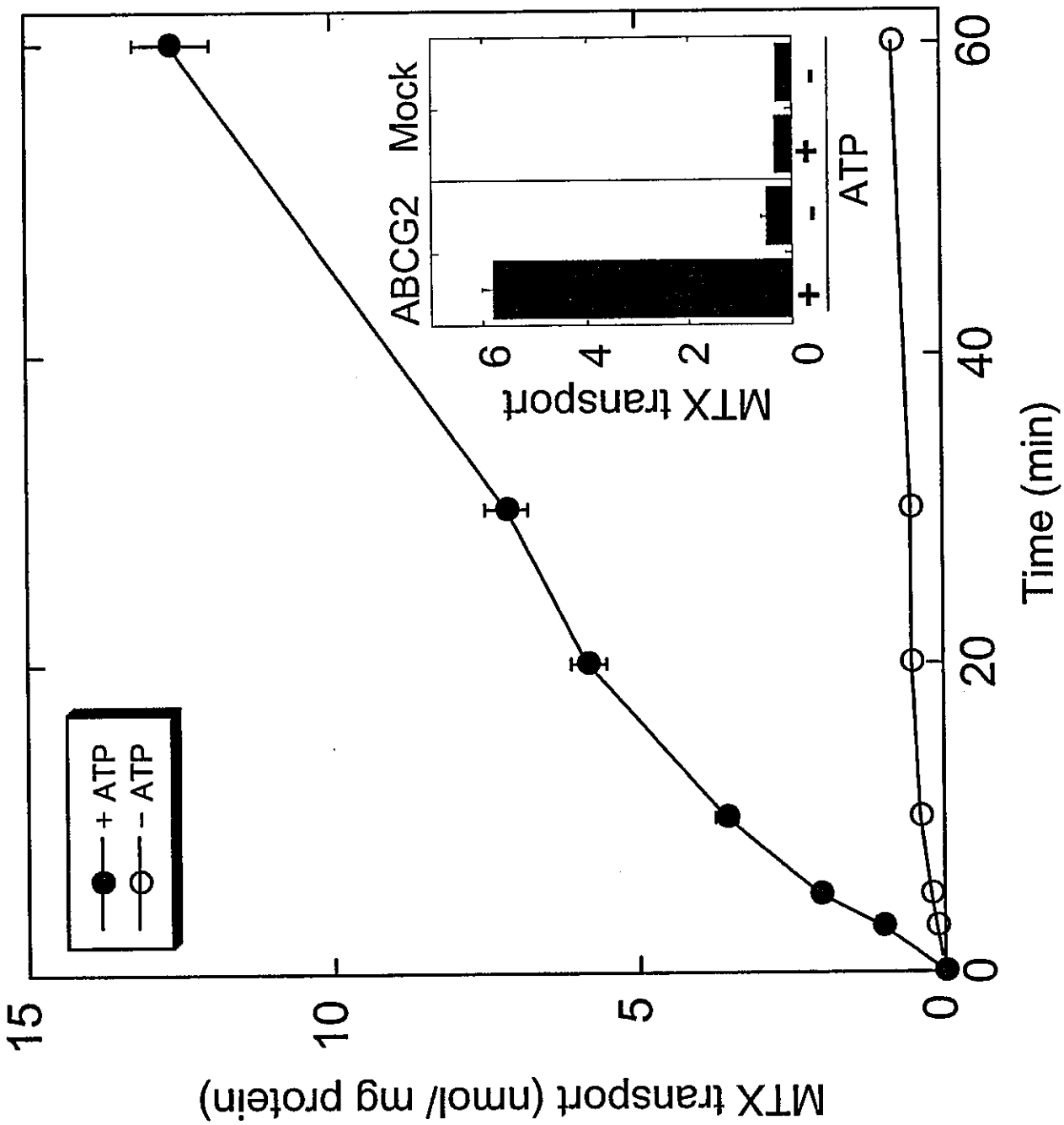


Fig. 6 Ishikawa et al.

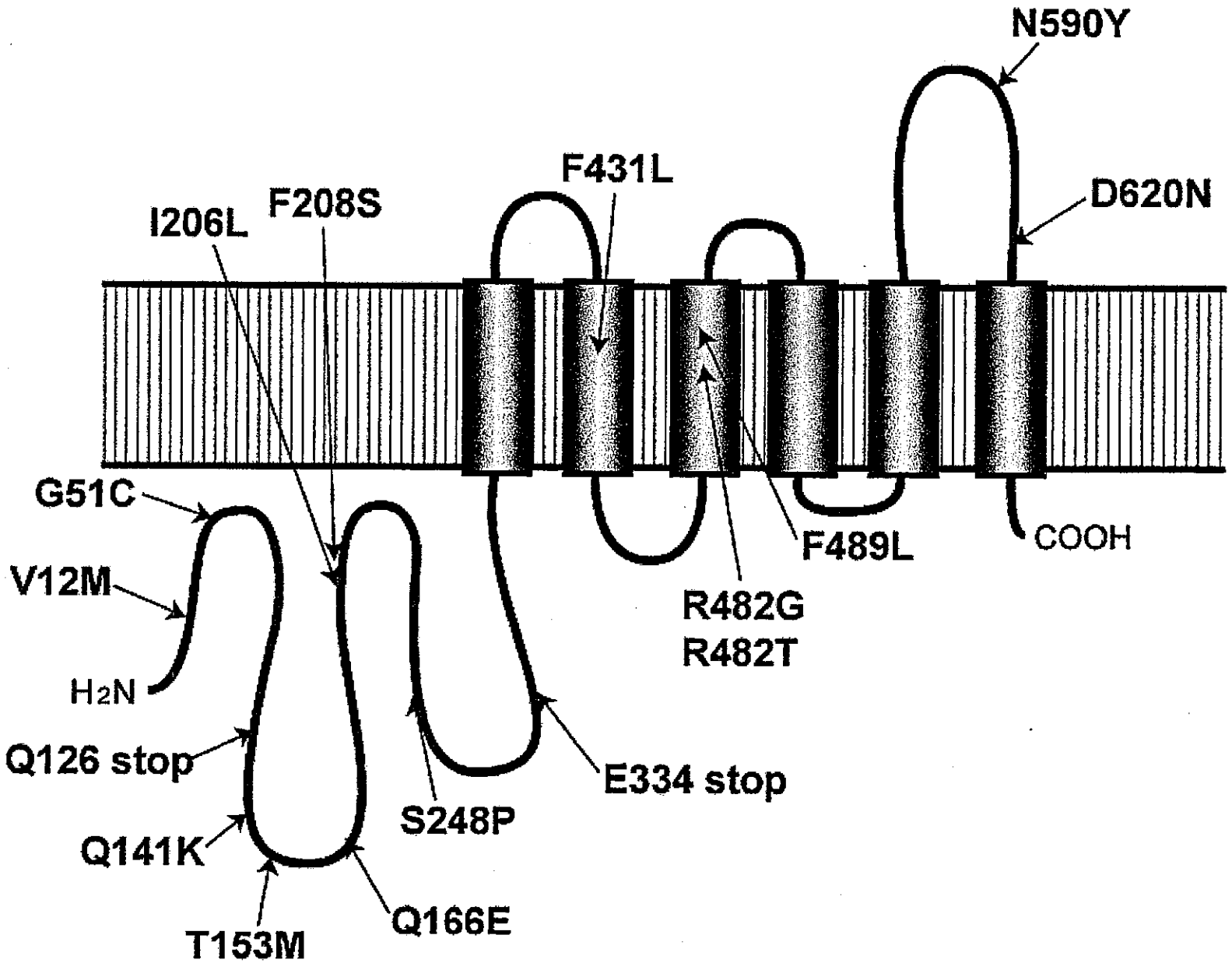


Fig. 7A Ishikawa et al.

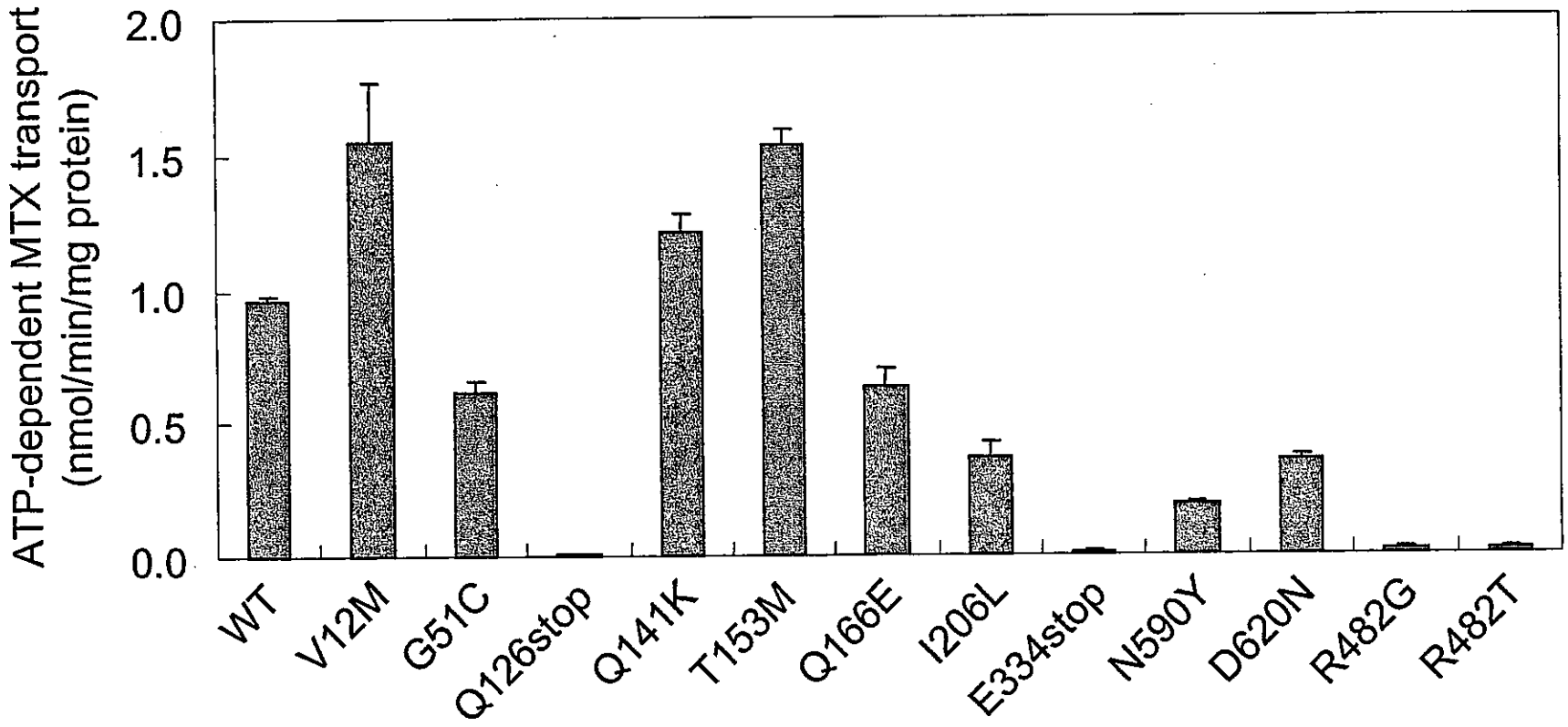


Fig. 7B Ishikawa et al.

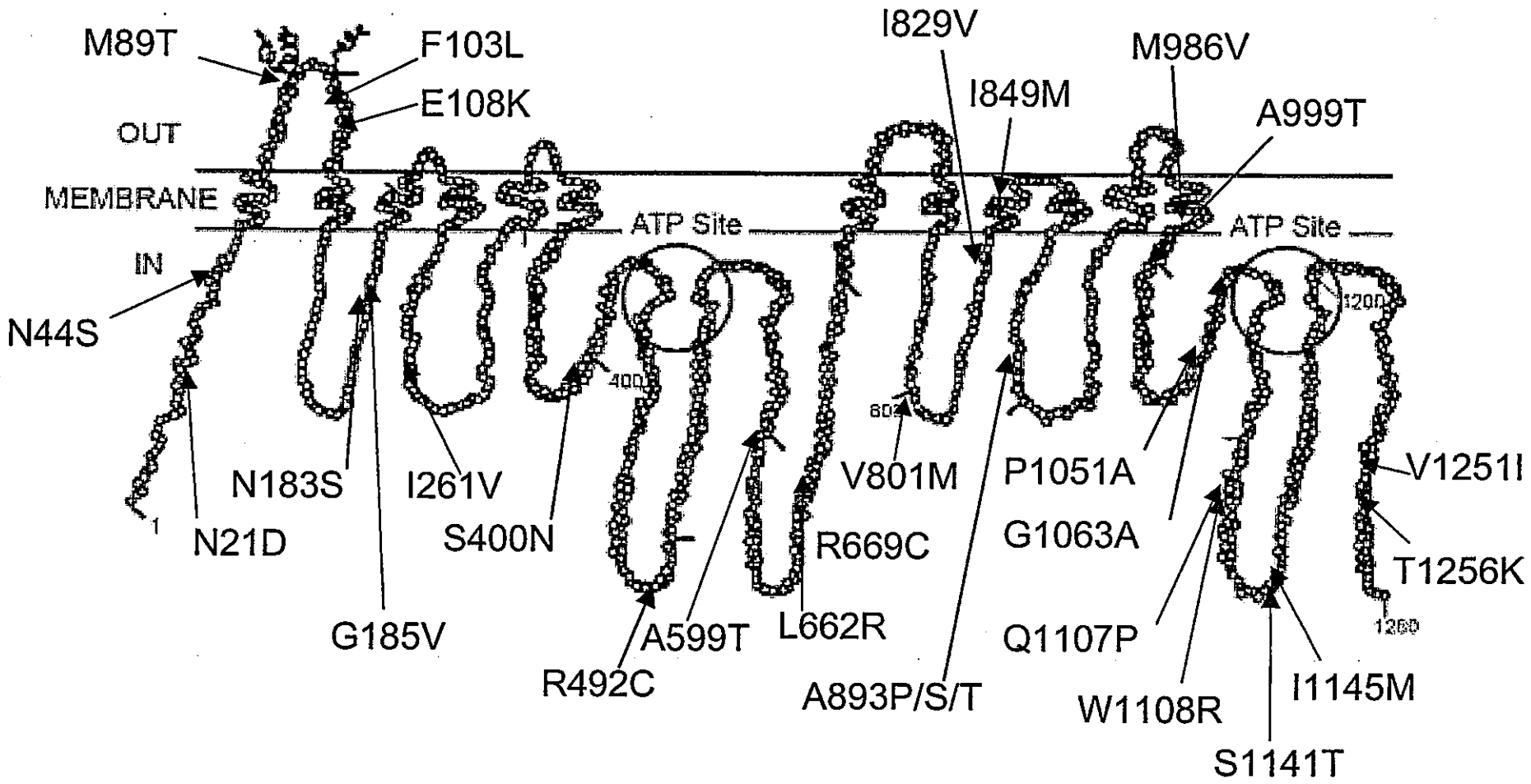
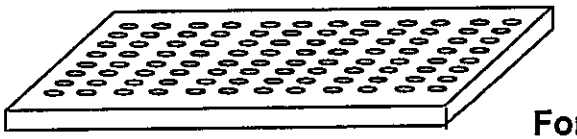


Fig. 8 Ishikawa et al.



For each well:

10 μl of cell membrane (2μg protein) in Incubation Buffer Solution

10 μl of test compound

Pre-incubation at 37℃ for 3 min

20 μl of 4 mM ATP/Mg in Incubation Buffer Solution

Incubation at 37℃ for 0 min or 30 min

20 μl of 5% TCA

42 μl of Solution A

18 μl of Solution B

Incubation at room temperature for 2 min

120 μl of Solution C

Incubation at room temperature for 1h

Measure Absorbance at 625 nm

Incubation Buffer Solution
50 mM Tris-Mes (pH 6.7)
2 mM EGTA
2 mM Ouabain
2 mM Dithiothreitol
50 mM KCl
5 mM NaN_3

Solution A
2N HCl : 0.1M Na_2MoO_4 = 4:3

Solution B
0.084%(w/v) Malachite Green
in 1% Polyvinyl Alcohol Solution

Solution C
7.8%(v/v) H_2SO_4

Fig. 9 Ishikawa et al.

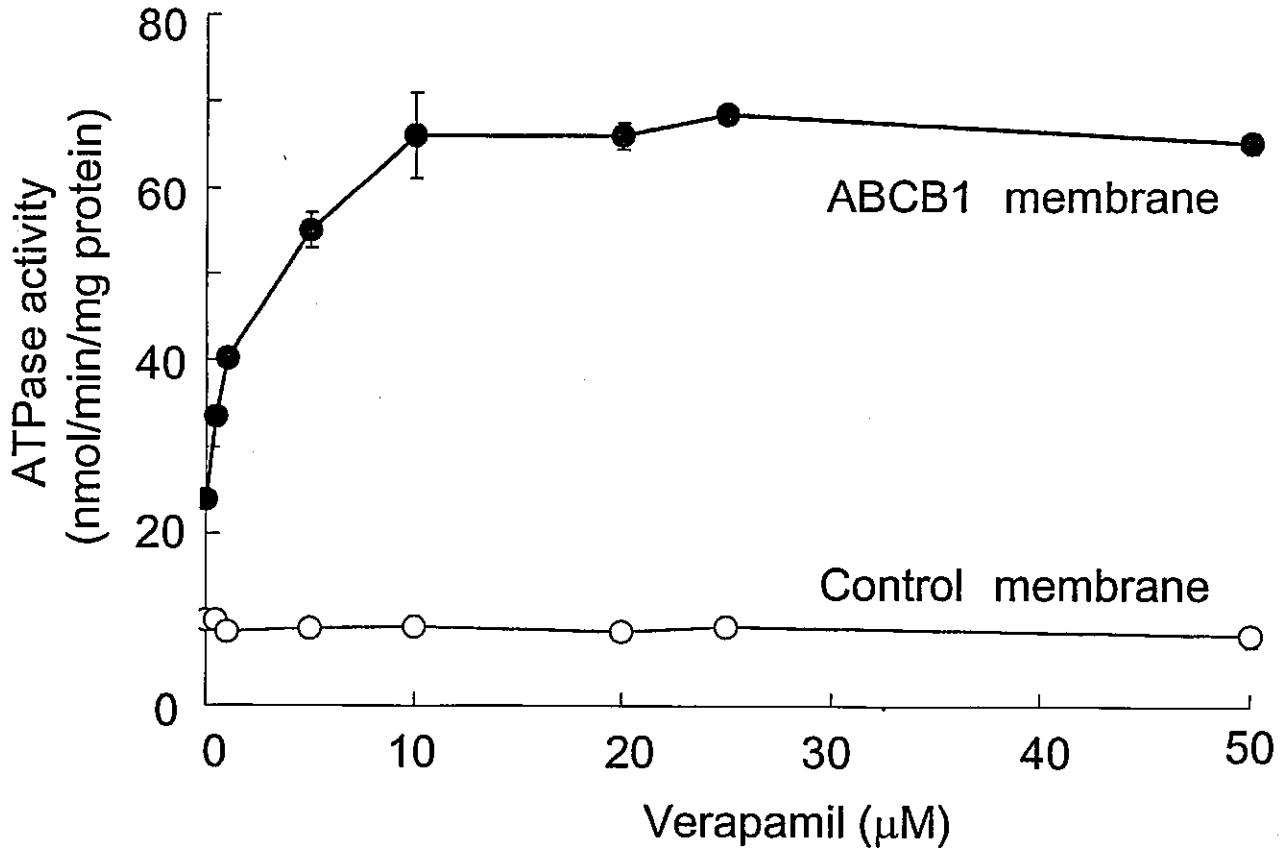


Fig. 10A Ishikawa et al.

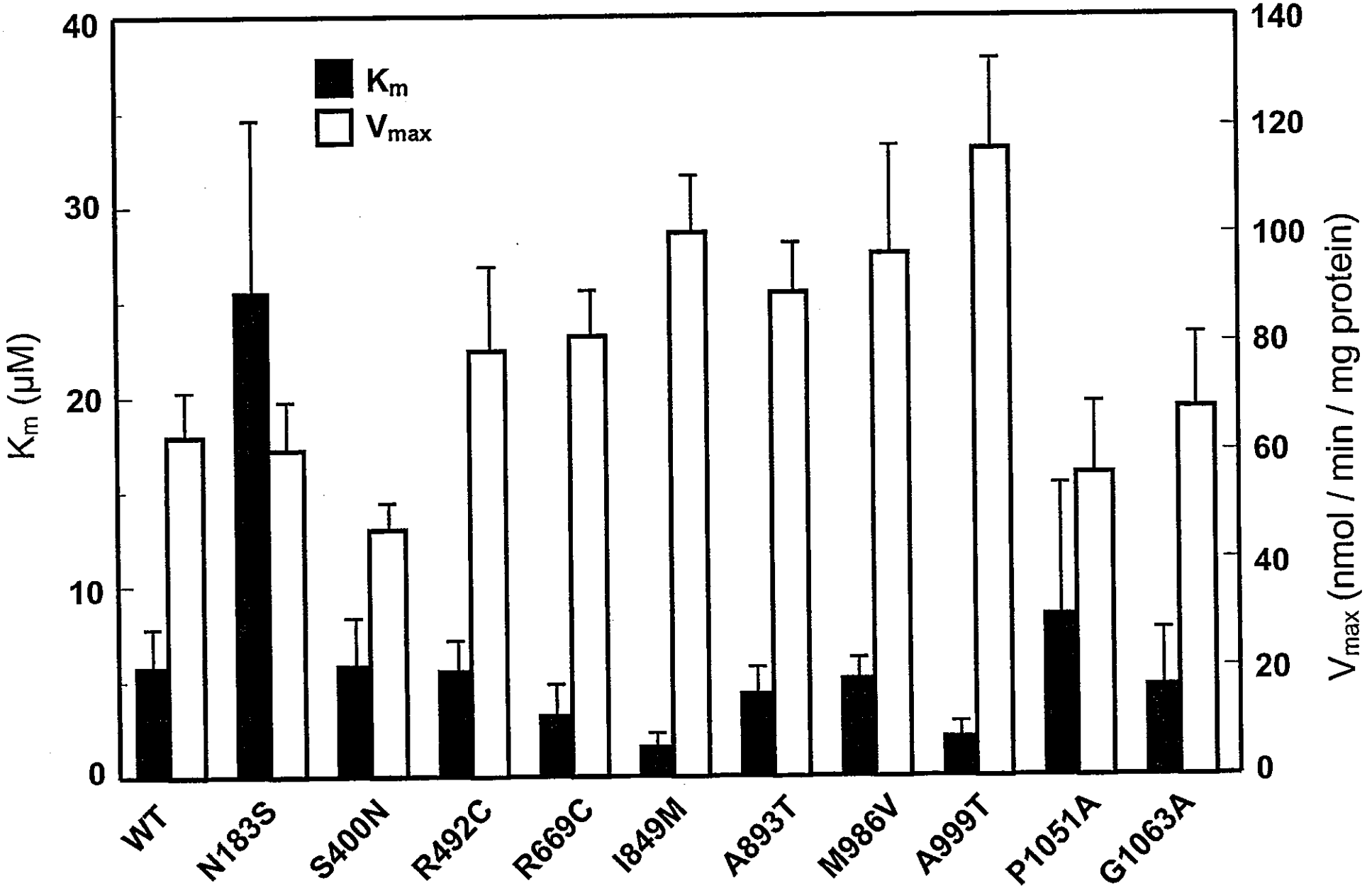


Fig. 10B Ishikawa et al.

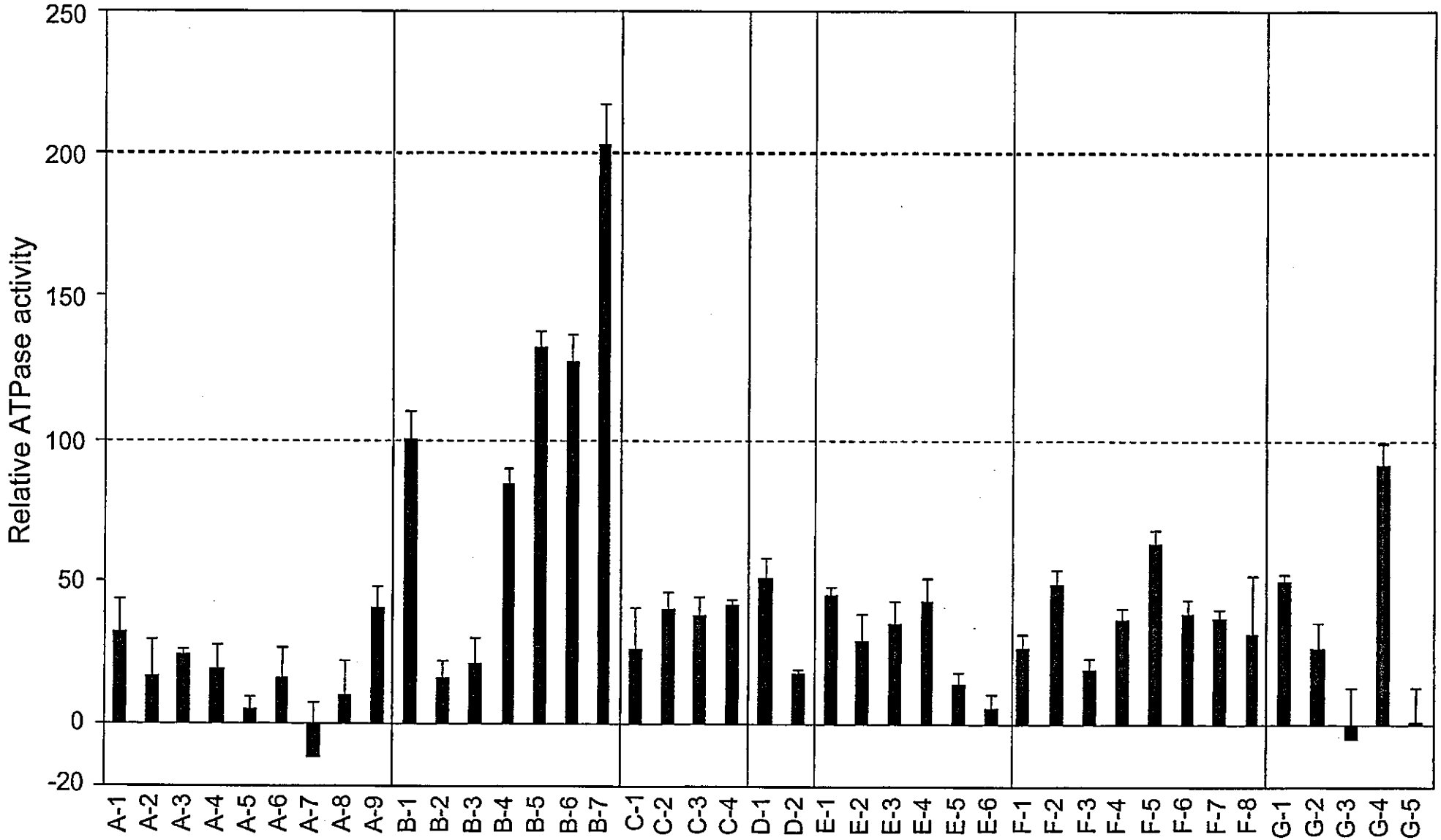


Fig. 11 Ishikawa et al.

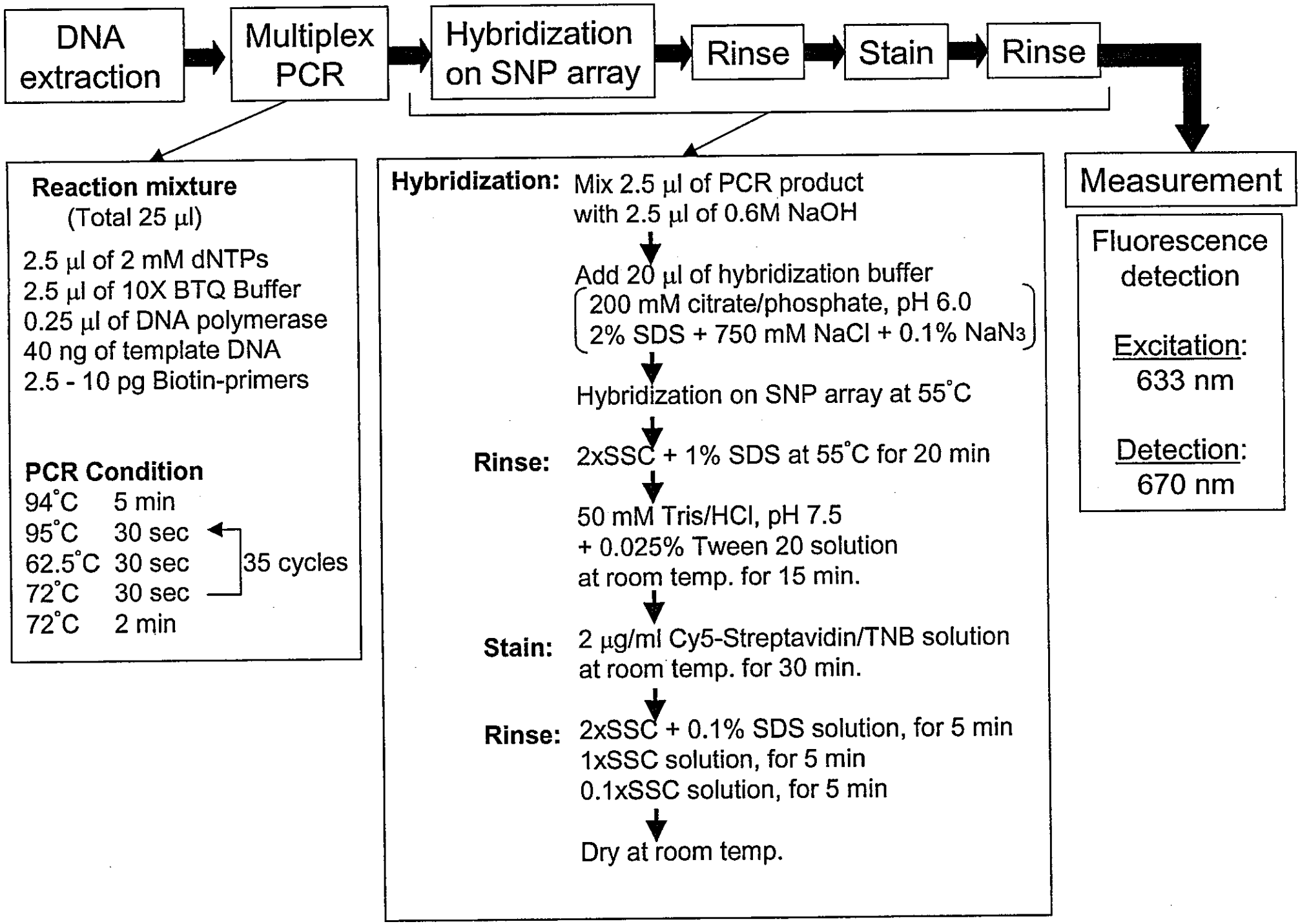


Fig. 12 Ishikawa et al.

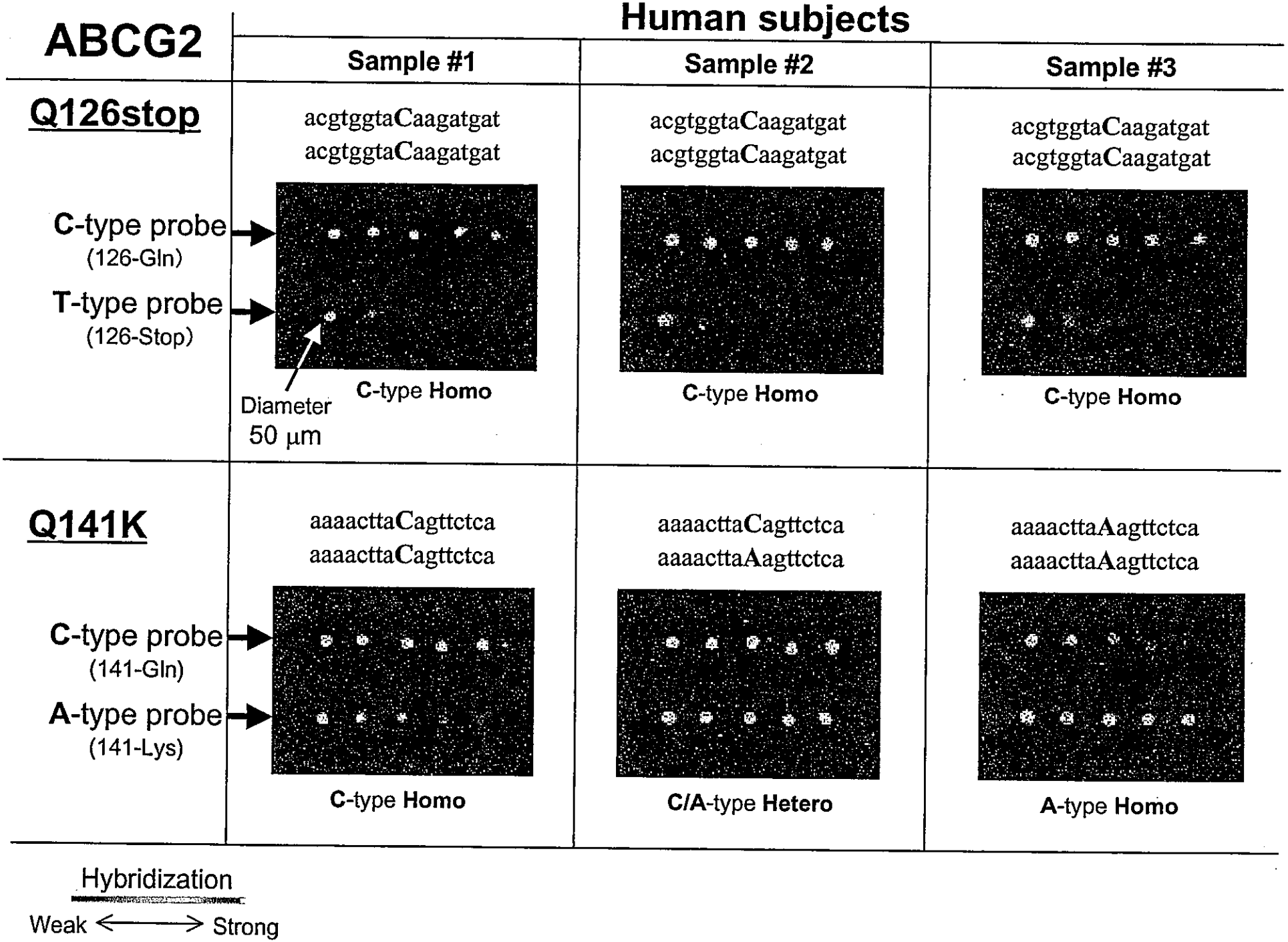


Fig. 13 Ishikawa et al.

Articles in PresS. Physiol Genomics (February 15, 2005). doi:10.1152/physiolgenomics.00107.2004

A meta-clustering analysis indicates distinct pattern alteration between two series of Gene Expression profiles for induced ischemic tolerance in rats

Makoto Kano[(1)], Shuichi Tsutsumi[(2)], Nobutaka Kawahara [(3)(4)], Yan Wang [(3)], Akitake Mukasa [(2)(3)], Takaaki Kirino [(3)(4)] and Hiroyuki Aburatani[(2)]

[(1)]Intelligent Cooperative System, Department of Information Systems, Research Center for Advanced Science and Technology, University of Tokyo, 153-8904, Japan
[(2)]Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan
[(3)]Department of Neurosurgery, Faculty of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[(4)]SORST (Solution-Oriented Research for Science and Technology) / JST(Japan Science and Technology), 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan

Running head: Visualization for Time-series Gene Expression Analysis

To whom requests for proofs are to be addressed:
Makoto Kano,
E-mail: mkano@cyber.rcast.u-tokyo.ac.jp

Abstract

We have developed a visualization methodology, called a *Cluster Overlap Distribution Map* (*CODM*), for comparing the clustering results of time-series gene expression profiles generated under two different conditions. Although various clustering algorithms for gene expression data have been proposed, there are few effective methods to compare clustering results for different conditions. Using *CODM*, the utilization of three-dimensional space and color allows intuitive visualization of changes in cluster set composition, changes in the expression patterns of genes between the two conditions, and relationship with other known gene information, such as transcription factors. We applied *CODM* to time-series gene expression profiles obtained from Rat 4-vessel occlusion models combined with systemic hypotension and time-matched sham control animals (with sham operation), identifying distinct pattern alteration between the two. Comparison of dynamic changes of time series gene expression levels under different conditions are important in various fields of gene expression profiling analysis, including toxicogenomics and pharmacogenomics. *CODM* will be valuable for various types of analyses within these fields since it integrates and simultaneously visualizes various types of information across clustering results.

Key words: time-series, transcription factor, visualization

1. Introduction

Advances in microarray technologies have made it possible to comprehensively measure gene expression profiles. Observation of dynamic changes of gene expression levels provides important markers to clarify cellular responses, differentiation, and genetic regulatory networks. In particular, a comparison of dynamic changes of time series gene expression levels under various conditions (e.g. administration of different drugs) is expected to make a major contribution to the understanding of complex biological processes. In general, we observe the influence of each condition through the results of clustering analysis, which is the most popular analysis for gene expression profiles. Therefore, a comparison between the results of clustering analyses in different conditions will allow interpretation of different macroscopic phenomenon that occurred under those conditions. However, although many clustering algorithms, including hierarchical clustering (1,2,4,15), k-nearest neighbor (17) and self-organizing maps (10,13,16) have been proposed, there are few effective methods to effectively compare clustering results under different conditions. We have defined four issues to be addressed for a comparison of clustering results, especially for a comparison of time series gene expression data under two different conditions: changes in the composition of the cluster sets, changes in the expression patterns, integration with known other gene information, and threshold problems.

Changes in the composition of the cluster sets

In this report, we focused on hierarchical clustering since it is the most popular method for gene expression analysis. Here we define the composition of a cluster set as the hierarchical structure of clustering results and "cluster set" as the set of all clusters in the structure. A comparison of clusters' compositions shows which clusters are conserved in different conditions and how the genes in a cluster for one condition are distributed into a cluster set under another condition. Genes that cluster under a single condition may possibly be regulated by the same factors for that condition. However, under different conditions, some of those genes would be regulated by other factors and generate different clusters. Thus, changes in the cluster compositions could provide key information for interpreting the

effects of the different conditions. To get a full picture of the relationships of two cluster sets, the overlap between each pair of clusters under the two different conditions should be evaluated. However, since clustering analysis, especially hierarchical clustering, almost always generates a great number of clusters, there are a very large number of combinations of clusters. Simple line connections of the genes between the dendrograms of two hierarchical clustering results (14) provides insufficient information about the relationships between the clusters. Therefore, an effective presentation method that provides a full picture of the relationships of the cluster sets would be desirable.

Recently, a statistical model for performing meta-analysis of independent microarray datasets was proposed (12). This model revealed, for example, that four prostate cancer gene expression datasets shared significantly similar results, independent of the method and technology used. However, in a comparison of the cluster sets based on different conditions, the objective is not to confirm that several datasets share significantly similar results, but to detect the differences between them. Several statistical algorithms have been proposed for evaluating how clusters based on expression profiles include genes with well-known functions (3,17). However, the number of clusters that were compared was limited and an effective presentation method was not required in those situations.

Changes in the expression pattern

Where two clusters under different conditions have a statistically meaningful number of genes in common, it is also important to examine the differences in their expression patterns. The differences of macroscopic phenomena that the conditions exhibit result from the differences of expression of multiple, rather than single, genes. Therefore, the genes whose expression patterns changed in a similar fashion between different conditions provide markers for the different phenomena. In other words, if the genes in a certain cluster based on one condition also constitute a cluster for another condition, but the expression patterns are greatly different between the two conditions, these genes are causally implicated in the phenotypic difference.

In general, there will be many false candidate genes whose expression patterns coincidentally match between the two different conditions. Therefore, it is important to simultaneously evaluate the statistical significance of the overlaps between clusters and the differences in their expression patterns.

Integration with other known gene information

In gene expression analysis, it is important to biologically interpret the results after integrating them with other known gene information. Therefore, changes in the composition of the cluster sets and changes in the expression patterns between different conditions should be associated with other known gene information such as transcription factors.

Threshold problems

In a comparison of cluster sets on gene expression profiles, we have to handle four types of thresholds: 1) a threshold for generating clusters for each condition; 2) a threshold for evaluating the number of common genes that two clusters have; 3) a threshold for evaluating the differences in the expression patterns between two clusters; and 4) a threshold for evaluating the relationship with other known gene information. Among these, determining the threshold for generating clusters is most challenging, because the clustering result strongly depends on this threshold, and a change of this threshold greatly affects the number and composition of clusters. It is generally difficult to determine optimal values for these four types of thresholds, and the results of analysis are greatly affected by the threshold values specified. Arbitrary selection of thresholds involves a risk of overlooking important genes, so the number of thresholds should be reduced and, if used, it is necessary to allow users to interactively change the thresholds.

We focused on visualization technology to address these four issues. Interactive visualization is effective for handling ambiguous threshold problems and for providing a wide variety of information at one time. In previous work, we developed a *Cluster Overlap Distribution Map* (*CODM*), which is a visualization method for comparing cluster sets based on different sets of gene expression profiles (7). In

this report, we extended it for time-series gene expression analysis. In the *CODM*, the relationships of all possible pairing of clusters can be examined and both the changes in the composition of the cluster sets and the changes in the expression patterns of the clusters can be effectively visualized as 3D histograms, without any arbitrary thresholds. In addition, relationships with other known gene information such as transcription factors can also be elucidated. We applied the *CODM* to a comparison between the gene expression datasets of double ischemia rats and sham control rats (with sham operation), and confirmed that *CODM* identified distinct patterns between the two.

CODM, available on our web site (http://www.genome.rcast.u-tokyo.ac.jp/CODM), runs on a PC with Windows 2000 or Windows XP. Memory requirement is in proportion to the square of the number of genes to be analyzed. The analysis for approximately 4000 genes, represented in this paper, required approximately 250 Mbytes. In addition, since the analysis results of the *CODM* are visualized by use of the OpenGL, a machine with a graphic board with a hardware accelerator for the OpenGL is recommended.

2. Materials and Methods

Experiment Design

In this report, *CODM* is illustrated using time-series gene expression datasets obtained from Rat 4-vessel occlusion models combined with systemic hypotension and time-matched control animals with sham operation. In the experiment, we used 2-minute ischemia rats with induced ischemic tolerance (*tolerant rats*: TOL) and rats with sham operation (*sham rats*: SHAM), after confirming the histological outcomes. Note that the sham rats did not acquire ischemic tolerance. Three days after the operation, we conducted a 6-minute ischemia operation on the two groups. Because of their ischemic tolerance, very little neuronal death of CA1 hippocampal neurons was observed in the tolerant rats (9). Using duplicate assessments of 6 time-points ({0h, 1h, 3h, 12h, 24h, 48h} x 2) after the second ischemia, microdissected CA1 regions from each of the two groups were subjected to oligonucleotide-based microarray analysis.

All animal-related procedures were conducted in accordance with guidelines for the care and use of laboratory animals set out by the National Institutes of Health and approved by the committee for the use of laboratory animals in the University of Tokyo. More detailed experimental design is described in our previous report (8).

Gene Chip experiment

Five μg of total RNA from each sample were used to synthesize biotin-labeled cRNA, which was then hybridized to a high-density oligonucleotide array (GeneChip Rat RG_U34A array, Affymetrix) essentially following a previously published protocol (6). The arrays contain probe sets for 8737 rat genes and ESTs, which were selected from Build 34 of the UniGene Database (derived form GenBank 107, dbEST/11-18-98). Sequences and GenBank accession numbers of all probe sets are available from the Affymetrix home page (http://www.affymetrix.com/index.affx.). Washing and staining was performed in a Fluidics Station 400 (Affymetrix) using the protocol EukGE-WS2. Scanning was performed on an Affymetrix GeneChip scanner to collect primary data. The Affymetrix Microarray Suite v4.0 was used to calculate the average difference for each gene probe on the array, which was shown as an intensity value of gene expression defined by Affymetrix using their algorithm. The average difference has been shown to quantitatively reflect the abundance of a particular mRNA molecule in a population (6). To allow comparison among multiple arrays, the average differences were normalized for each array by assigning the mean of overall average difference values to be 100. This dataset has been submitted as GSE1357 to the National Center for Biotechnology Information's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/info/linking.html)

Preprocessing and clustering

In the following analysis, we used datasets as 12 time-point ({0a, 0b, 1a, 1b, 3a, 3b, ..., 48a, 48b} = $\{T_i\}$ (i = 1,2,...,12)) datasets on TOL and SHAM, since the *CODM* does not depend on the intervals of the time-points.

Standard clustering analysis for gene expression profiles is based on the correlation coefficients between genes. Therefore, this approach can not handle genes with expression profiles that have almost no changes for a condition. However, if the expression profiles of those genes have meaningful changes in expression levels for other conditions, they provide markers to interpret the influence that the conditions exerted, because they are possibly regulated by different factors. To handle those genes and to align the baselines of the expression patterns between the different datasets, preprocessing (i.e. filtering and normalization) was conducted for all of the datasets where TOL and SHAM were merged. More specifically, 3,363 probes with mean expressions above 50 and coefficient of variance (=standard deviation / mean) above 0.1 were selected. After logarithmic transformation of the gene expression data, the expression levels were normalized to satisfy the following equations:

$$\sum_{i}^{12}(x_i + y_i) = 0 \quad (1)$$

$$\sum_{i}^{12}(x_i^2 + y_i^2) = 1 \quad (2)$$

where x_i and y_i are normalized expression levels of a gene at time-point T_i (i = 1,2,...12) on conditions TOL and SHAM, respectively. Using these normalized datasets, hierarchical clustering analysis based on Euclidian distances was then performed for each dataset independently. Clustering analysis using Euclidian distances instead of correlation coefficients allows us to handle genes whose expression levels are down-regulated or up-regulated. In addition, due to the common normalization, gene expression patterns can be compared within a dataset and between datasets.

In general, Euclidian-distance based clustering after normalization, in terms of mean and standard deviation, is equivalent with correlation-coefficient based clustering. That is, a Euclidian-distance based clustering analysis for the merged data of TOL and SHAM with the above preprocessing is equivalent with a correlation-coefficient based clustering analysis for the original merged data. In the analysis of the *CODM*, the preprocessing is conducted for the merged data, and Euclidian-based clustering is individually conducted for each data. Roughly speaking, this analysis provides us with results similar to

those of normal correlation-coefficient based clustering, while it allows us to handle genes with expression profiles that have changes for only one condition but not for the other.

As Figures 1a and 1b show, there are a large number of clusters generated at various levels. Although the composition and number of cluster sets depend on the threshold value of the distance, it is generally difficult to identify an optimum value. These aspects make it difficult to compare cluster sets derived from different sources.

The cluster overlap distribution map (*CODM*)

The *CODM* is a visualization methodology for pair-wise comparison between cluster sets generated from different gene expression datasets. In this methodology, two types of cluster sets (i.e. dendrograms of hierarchical clustering results) are mapped respectively to the X-axis and on the Y-axis, and the relationship between them is displayed as a 3D histogram (Figure 2). In this report, the dendrogram of TOL is mapped to the X-axis and that of SHAM is mapped to the Y-axis. The statistical evaluation values of the overlaps between two clusters selected from the respective cluster sets are displayed as the height of the blocks (Figure 2). More specifically, we evaluated the number of common genes between the two different clusters by using hypergeometric probability distributions (17). Assuming that the generation of gene clusters is a random selection from among the total set of genes, the probability of observing at least (k) overlapping genes between randomly selected (n_1) genes and (n_2) genes from among all of the (g) genes is given by:

$$P(g,n_1,n_2,k) = 1 - \sum_{i=k}^{k-1} \frac{{}_{n_2}C_i \cdot {}_{g-n_2}C_{n_1-i}}{{}_{g}C_{n_1}} \quad \left(= P(g,n_2,n_1,k)\right) \qquad \textbf{(3)}$$

When the P-value is small, the overlap is regarded as statistically meaningful. Thus, we defined the evaluation value of the overlap as:

$$E(g,n_1,n_2,k) = -\log_{10} P(g,n_1,n_2,k) \qquad \textbf{(4)}$$

Then in the area (R_{ij}) determined by a cluster on the X-axis (X_i) and a cluster on the Y-axis (Y_j), a block whose height represents $E(g, n_{xi}, n_{yj}, k_{ij})$ is displayed, where (n_{xi}) is the number of genes in (X_i), (n_{yj}) is the number of genes in (Y_j), and (k_{ij}) is the number of overlapping genes between (X_i) and (Y_j) (Figure 2). We term this block an *overlap block*. Note that the number of UniGenes, to which probes in a cluster correspond through their original GenBank accession number, was used as the number of genes. In this report, all 8737 probes on RG-U34A were corresponding to 5,249 UniGenes ($g = 5,249$).

For hierarchical clustering, there are a large number of clusters generated at various distance levels. Our algorithm examines the overlaps of the genes between all combinations of two clusters with smaller *distance level* values than the *cut level*, which is a threshold value specified by users (Figure 1). In other words, we evaluated and visualized any clusters with a smaller distance level than the *cut level*, even if they were included in other clusters. Note that conventional hierarchical clustering does not focus on sub-clusters that are included in other clusters. Since all of the statistically significant combinations between cluster sets can be visualized simultaneously, users can grasp the overall picture of the relationships between the two different cluster sets.

In the *CODM*, all of the clusters are dealt with equally without regard to their difference level (i.e. their homogeneity). Even if they are included in other clusters, all of the statistical significance of the number of common genes between clusters is simultaneously visualized. Therefore, there is a risk that a small *overlap block* may be hidden by a large block. For example, assume that the clusters X_j and Y_n are included in X_i and Y_m respectively. Then, if the evaluation value E_{jn} is less than E_{im}, the small block B_{jn} will be hidden in the large block B_{im} (Figure 3a). To avoid this problem, the *CODM* allows the user to change the *cut level* interactively. That is, if the user decreases the *cut level*, some small blocks that are hidden in larger blocks will emerge. Therefore, in consideration of the homogeneity of clusters and the relationships with other gene information, the user can find important genes displayed as blocks in the *CODM*.