System of Assisting in Discovery by Designing Views:
Data and Experiment

Osamu Maruyama Tomoyuki Uchida Satoru Miyano

1東京大学 医科学研究所  2広島市立大学 情報科学部
I.M.S., University of Tokyo F.I.S., Hiroshima City University

Abstract: We report some of hypotheses generated in a series of computational experiments
on data of scientific domains with the computational system HYPOTHESISCREATOR, which is
designed to assist researchers and experts in the process of discovery. One of the features
of HYPOTHESISCREATOR is that the system allows the user to design views on data. The
hypotheses are constructed with several views good for explaining given data, which are selected
among over ten millions designed views.

1 Introduction

There are discussions on the roles of the developers
and users of discovery programs. Langley [6] identified
five steps during which developers or users can influ-
ence the behavior of a computational discovery sys-
tem, and strongly recommended that computational
systems provide more explicit support for human in-
tervention in the discovery process. De John recog-
nized the important roles of developers of discovery
programs, and proposed guidelines for developers on
achieving the integration of their tools in so-called dis-
cover environments [4].

We agree with Langley’s recommendation. In fact,
in [9], we have claimed the importance of views on data
in the process of discovery, and proposed the system
GENOMICHYPOTHESISCREATOR that allows the users
to design their own views on data. The key concepts
of the system are viewpoints and views. Informally,
a view is a pair of a polynomial-time algorithm of inter-
preting data and a set of parameters of the algorithms
called viewpoints, and a view is a pair of
such an interpreter and a fixed viewpoint.

We are developing a new version of the system called
HYPOTHESISCREATOR. In the new version, a hierar-
chical clustering program is available, and new view-
scopes for numerical table data are added. We can in-
voke a clustering program AUTOCLASS [2] as an exter-
nal hypothesis generator from HYPOTHESISCREATOR.
The aim of this extension is to make clusters of genes
by applying such views to expression profiles pro-
duced by microarrays. In addition, viewpoints are
modified to deal with data in Japanese codes for ex-
traction rules of traditional Japanese poetry, WAKA.

An outline of the system is as follows: The user can
conduct the following processes: (i) collecting data
from databases, (ii) designing views and viewscapes on
the data, (iii) selecting a hypothesis generator whose
input is the resulting values of the interpretations of
viewscapes on the data, (iv) selecting a strategy of
search for good views. After these processes, the sys-
tem generates a hypothesis for a particular viewscope
by the selected hypothesis generator, and determines
the next view to be examined along with the selected
search strategy from the evaluation value of the re-
sulting hypothesis and the viewscope. The system re-
peats this process until the termination condition of
the search strategy is satisfied.

In this paper, we report some of hypotheses gener-
ated in a series of computational experiments on data
of scientific domains with the computational system
HYPOTHESISCREATOR. We have used the system in
SUN ENTERPRISE 450 of 4 CPU’s, 3000 of 6 CPU’s,
3500 of 8 CPU’s, and 10000 of 64 CPU’s. In these
experiments, we have designed over ten millions views
and applied them to genomic data. Actually, some
of the hypotheses we have obtained show well known
features of data that should be explained. Our exper-
iments are still carried out for various data. We will
report the results in the full version paper.

2 Preliminaries

We define the key concepts, a view and a viewscope,
in the following way:

Definition 1 Let \( \Sigma \) be a finite alphabet. A viewscope
over sequences on \( \Sigma \) is a pair \( V = (\psi, P) \) of an al-
gorithm \( \psi \), called the interpreter of \( V \), with two input
sequences over \( \Sigma^* \times \Gamma^* \), where \( \Gamma \) is a finite alphabet and
\( P \subseteq \Gamma^* \) is a set satisfying the following conditions:
1. For $x \in \Sigma^*$ and $y \in \Gamma^*$, $\psi$ on $(x, y)$ outputs a value $\psi(x, y)$ in a set $W$ called the value set of $V$ if $y \in P$ and "undefined" otherwise.

2. For $x \in \Sigma^*$ and $y \in \Gamma^*$, $\psi$ on $(x, y)$ runs in polynomial time with respect to $|x|$ and $|y|$.

An element in $P$ is called a viewpoint of $V$. For a fixed viewpoint $y' \in P$, the function $\psi_{y'}(x) = \psi(x, y')$ is called a view from the viewpoint $y'$ over sequences on $\Sigma$.

For a set $D \subseteq \Sigma^*$ and a viewscope $V = (\psi, P)$ over sequences on $\Sigma$, we call the $|D| \times |P|$ matrix $D^V$ defined by $D^V(x, y) = \psi(x, y)$ for $x \in D$ and $y \in P$ the data matrix of $D$ under the viewscope $V$.

## 3 Data and Experiments

We describe our computational experiments on genomic data with HYPOTHESISCREATOR. For these experiments, we have selected, as data, the annotation files of the complete genome sequences of Saccharomyces cerevisiae, which can be down-loaded from the ftp site [14], whose total size is 43MBytes. These files are formatted in the form of DDBJ/EMBL/GenBank feature table, each of which consists of three types of objects: the header part, the chromosome sequence, and its annotations. Saccharomyces cerevisiae is a kind of baker’s yeasts, which has 16 chromosomes whose DNA sequences are $12,057,849$ bp in total, and which codes $6,565$ genes including putative ones. Saccharomyces cerevisiae has been extensively studied, so that it is often used as a model organism for various researches in biology.

Note that it is possible for other organisms to carry out the following experiments if their annotated complete genome sequences are provided. Actually, we are going to such process, whose result can be shown elsewhere.

As a work related to a knowledge discovery from yeast genome sequences, there is a work of Brazma et al. [1] which reported a result of discovering transcription factor binding sites from the upstream regions of genes in the genome sequences by a method they developed to the highest rating patterns in given sequences. The rate of rating patterns adopted in their discovery strategy can be thought as a view on sequences. Their system would be more useful if other views are available in the system. In the second example in this paper of our experiments, we have intensively designed viewscopes on the upstream regions of genes, and shown characteristic patterns appearing in the obtained hypothesis.

### 3.1 Cyclin Genes

Here we describe an experiment to generate knowledge explaining cyclin genes of Saccharomyces cerevisiae, which appear to play roles of switches in cell cycles. The somatic cell cycle is the period between two mitotic divisions. The time from the end of one mitosis to the start of the next is called interphase, which is divided into the G1, S, G2 periods. A cyclin accumulates by continuous synthesis during interphase, but is destroyed during mitosis. Its destruction is responsible for inactivating M phase kinase and releasing the daughter cells to leave mitosis (see [7]).

The aim of this experiment is to extract some knowledge of cyclin genes from the annotated complete genome sequences of Saccharomyces cerevisiae.

We have a list of cyclin genes as a search result of Yeast Proteome Database [17]. Using the list, HYPOTHESISCREATOR separates the genes described in the annotation files into cyclin genes and the other genes. The number of cyclin genes is 18 and that of the other genes is 6565 including RNAs.

We have designed various kinds of viewscopes on the DNA and amino acid protein sequences as follows: approximate string matching viewscopes on DNA sequences, approximate string matching viewscopes on amino acid sequences, and PROSITE viewscopes, where PROSITE [19] is a database of protein families and domains which classified with a kind of regular expressions. Note that a mismatch has three types: insertion, deletion, and substitution. In the approximate string matching viewscopes, any combination is accepted. In fact, the hypothesis in Fig. 3 is generated through approximate string matching views with insertion and deletion. In approximate string matching views, we can specify the threshold of the number of occurrences of patterns. The default of the threshold is one. As an option, these viewscopes can be filtered with alphabet-indexing [10], which is installed in knowledge discovery system BONSAI[11].

By using these designed viewscopes, we have conducted HYPOTHESISCREATOR, which has produced distinctive hypotheses. One of them is given in Fig. 3. To obtain the hypothesis, we have used 3,305 approximate string matching viewscopes with insertion and deletion on amino acid sequences. The patterns are automatically prepared by extracting from text regions with the specific lengths. The total number of viewscopes, in this case, corresponding to the number of patterns, is 935,230. We have repeated such an experiment over 20 times with different viewscopes, whose total number of viewscopes is 13,631,690.

We can see that 18 cyclin genes are roughly classified into 3 groups. The view assigned to the root node of the decision tree in Fig. 3 is the approximate string matching view with pattern LRRISKAD of the allowed mismatch 4 such that the location of the text of a gene $g$ is the region from 225 to 375 of the translation of $g$, that is, an amino acid protein sequence. The CLB subfamilies, CLB1, CLB2, CLB3, CLB4, CLB5 and CLB6 of cyclin genes are completely separated from the other genes by the root node’s view. CLB1, CLB2, CLB3 and CLB4 are known as G2/M-phase-specific cyclins, and CLB5 and CLB6 are B-type cyclins accumulating late in G1. In Yeast Proteome Database [17], we can easily find the fact that pairwise identities and similarities of the 6 translations of the 6 CLB genes are relatively high. Furthermore, we can find that one of the most common subsequences among the 6 translations is the pattern LRRISKAD.

Among the genes $g$ not matched with the view of the root node, CLN1, CLN2, PCL1 and PCL2 satisfy the following rules: the pattern CLILAAK is matched with the mismatch at most two in the interval [90,150] of the translation of $g$, and the pattern KSN is matched with the mismatch at most one in the interval [0,50] of $g$. CLN1, CLN2 and PCL1 are known to be G1/S-specific cyclins.

For the another group of 6 genes which reach the deepest leaf of the decision tree in Fig. 3, we have not characterized them yet.
Fig. 1: Hypothesis for cyclin. Approximate string matching views with insertion and deletion are assigned to non-terminal nodes. The view \( v \) assigned to the root node is interpreted as follows: The pattern of \( v \) is LRRISKAD, and the number of allowed mismatches are at most 4. For each gene \( g \), the approximate string matching of \( v \) is applied to the subsequence from 225 to 375 of the translation of \( g \). The gene \( g \) flows into the sibling started with “+1-” if the pattern is matched in the subsequence with at most 4 mismatches and flows into the sibling with “+0-” otherwise. The 18 genes that should be explained and the other 6565 genes are reached to the node. For other internal nodes, we can interpret in the same way.

From this experiment, we recognize that we can obtain some of characterizations of data by applying volumes of views to data without expert knowledge.

3.2 DNA Replication Genes in late G1

The group of genes we next use is the genes reported as genes related to DNA replication in late G1 in [3], which are the following: RNR1, DUT1, DPB3, RFA3, RFA2, POL2, DPB2, CDC9, RFA1, PRI2, CDC45, CDC17, CDC21, POL12 and POL30.

Among our hypothesis creation on the genes, we report an experiment to characterize the genes in their upstream regions. In the experiment, we have designed 193 approximate string matching views on DNA sequences where the total number of viewpoints is 89,225. Fig. 3 is a decision tree representing the hypothesis produced by HypothesisCreator for the genes.

We can see that, in the created hypothesis, patterns of gc rich are a key. 7 genes (RFA1, PRI2, CDC45, CDC17, CDC21, POL12 and POL30) among the 15 genes to be characterized satisfy the following rules: Let \( g \) be a gene. The pattern “acgcgt” exists in the upstream region [-200,-150] of \( g \) (rule 1), the pattern “taccat” matches in [-140,-90] of \( g \) with at most one mismatch (rule 2), and the pattern “cactat” matches in [-110,-60] of \( g \) with at most one mismatch (rule 3). By this rule, the 7 genes, which are related to DNA replication in late G1, are completely separated from the other genes. On the other hand, 5 of the 8 genes not satisfying the rule 1 have the substring “actct” of the pattern “acgcgt” of rule 1 in the upstream regions [-130,-100]. We can easily see that 12 genes of 15 have such features.

TRANSFAC [18] is a database of transcription factors. We can find that the sequence “acgcgt” assigned to the root node is a transcription factor binding site of CDC21 [8, 12]. In addition, the pattern “acgcg” is a substring of a transcription factor binding site of CDC9. The first and last positions of factor binding sites of the transcription factor are -160 and -105, which is overlapped with the text region [-130,-100] of the approximate string matching view. This fact would be one of evidences of the effectiveness of HypothesisCreator.

Concluding Remarks

As a built-in hypothesis generator, a hierarchical clustering program is available in the current version of HypothesisCreator. In the clustering program, several similarity and distance functions between objects and clusters are selectable. Additionally, we can call AutoClass [2] as an external hypothesis generator from HypothesisCreator. To use these cluster generators effectively, we have implemented various kinds of views, through which the users can operate data and transform the data into different ones according to their own aims. Currently, we are making computational experiments to cluster genes with using gene expression profiles produced by microarrays, which are available at [15, 16]. We will report the results of clustering genes in a full paper version.

The phenomena of natively unfolded or natively disordered proteins have been found to have significant biological functions [13, 5]. HypothesisCreator is now being applied to classify these groups of disordered proteins in a more systematic and accessible manner for future data retrieval.

We have modified the views of Hypothet-
acgcgt(0) +0->acgcg(0) +0->tattacgc(0) +0->cggtcgtaaa(1)+0->NO (0,6409)
[-200,-150] | [-130,-100] | [-110,-70] | [10,30] |
(15,6534) | (8,6492) | (3,6409) | (1,6409) +1->YES (1,0)
| | | |
| +1->YES (2,0)
| |
| |

DPB3,DUT1
+1->tgtccat(1) +0->NO (0,75)
[-200,-150] |
(5,83) +1->aacgaa(1) +0->YES (5,0)
[-100,-50] | RFA3,RFA2,POL2,DPB2,CDC9
(5,8) |
| +1->NO (0,8)
|
+1->taccat(1) +0->NO (0,34)
[-140,-90] |
(7,42) +1->cactat(1) +0->NO (0,8)
[-110,-60] |
(7,8) |
+1->YES (7,0)
|
RFA1,PRI2,CDC45,CDC17,CDC21,POL12,POL30

Fig. 2: Hypothesis for DNA replication in late G1. Approximate string matching views with insertion, deletion, and substitution are assigned to non-terminal nodes. For example, the view assigned to the root node is the approximate string matching of the pattern “acgcgt” without any mismatch whose text region is the region from -200 to -150 of the upstream of each gene.

References


