Integration of Diverse Knowledge and Data into Biomedical Knowledge Matrices

Yoshimasa Tsuruoka[1] **Teruyoshi Hishiki**[2] **Osamu Ogasawara**[3] **Kousaku Okubo**[4]
[1]CREST, JST (Japan Science and Technology Corporation)
tsuruoka@is.s.u-tokyo.ac.jp
[2]Biological Information Research Center,
National Institute of Advanced Industrial Science and Technology (AIST)
t-hishiki@jbirc.aist.go.jp
[3]Information and Mathematical Science Laboratory
oogasawa@v001.vaio.ne.jp
[4]National Institute of Genetics
kousaku@genomatrix.com

Abstract

After the accomplishment of human draft sequence, more and more efforts are being made in the mapping of the data-driven patterns to background knowledge, hoping to efficiently produce hypotheses out of the flood of data. Here we propose a framework of biomedical data and knowledge that has a high adaptability to the automated data interpretation. Then, we show that biomedical databases with heterogeneous scopes and structures can be converted to the format, and possible roles of ontology of biomedical objects combined with natural language processing techniques. Lastly, we present applications of formatted biomedical knowledge to scientific discovery.

1 Introduction

After the accomplishment of human draft genome sequence (Lander, E. S. et al., 2001), systematic prediction of gene functions (Marcotte et al., 1999) is one of the major goals in biomedicine. Toward this goal, high-throughput measurement of gene (protein) features such as expression profiles (Iyer, V. R. et al., 99; Velculescu, V. E. et al., 1999; Hsiao, L. L. et al., 2001; Su, 2001) and protein-protein interactions (Ito, T. et al., 2001; Ho, Y. et al., 2002) has become a trend, and data is generated at an unprecedented rate. It is clear that hypothesis creation on the roles played by genes with systematic and integrative approaches (Scherf, U. et al., 2000; Greenbaum, D. et al., 2001) to collected data is anticipated as the next goal.

The first step toward that goal was the representation of measurement data, the extraction of global patterns from the data (Eisen, M. B. et al., 1998; Ge et al., 2001; Bussemaker et al., 2001; Rives and Galitski, 2003), and the visualization of the results (Gilbert et al., 2000). Here we summarize the measurement data into only two formats. One is the 'feature array', or an array of features and their values for a type of biological object, e.g. genes and cells, and the matrix as a collection of the arrays. For example, the structural database (Apweiler, R. et al., 2001) is a collection of genes with features such as protein families, domains, and functional sites. Another example is the expression profiles data set, which is a matrix of genes and tissues with each cell representing relative or absolute abundance of cognate transcripts. Combinations of two types of measurement data by making the product of the matrices can give a prediction of a new type of relation (Scherf, U. et al., 2000). The other format is the 'gene-gene correlation/similarity matrix' representing the strength of relations in all-to-all gene pairs. Some of the methods produce directly this type of data; on the other hand, the gene 'feature arrays' can be transformed to this type of data by calculating the correlations between the feature values of all-to-all gene pairs.

Now, more and more efforts are being made in the second step, or mapping the data-driven patterns to the integrated background knowledge e.g. biochemistry, cell biology, pathology, and medicine.

A problem here is to allocate authentic features to biological objects, and the efforts to work collaboratively to build controlled and classified vocabularies (Ogata, H. et al., 1999; Ashburner, M. et al., 2000) for molecular localization, actions, and roles, and then to assign them to genes (Xie, H., 2002; Camon, E. et al., 2003) are being promoted. Some of the controlled vocabularies are called 'ontology' because they have manually edited relations representing e.g. 'Is-a' and 'Part-of' relations, connecting the objects in a tree or network structure. By extending the scope of biological ontologies, e.g. relations between anatomical structures and tissues constituting them as represented in TissueDB (http://tissuedb.ontology.ims.u-tokyo.ac.jp), more diverse problems like matching tissues between expression profiling platforms will become easier.

We argue that there is another aspect of the integration problem: the format of data and knowledge to be interrelated with each other. The network representation of the relations between objects is widely used, and an experiment (Jenssen et al., 2001) extracted the networks between related genes using the co-occurrence frequency of gene symbols in MEDLINE articles. However, no study using such representation has successfully combined data and knowledge into a global view of a new type of relations as presented in the combination of genome-wide measurement data (Scherf, U. et al., 2000). What we will propose as the common data format will enable calculating the knowledge as well as the genome-wide measurement data, and will integrate the data and knowledge toward a discovery.

2 Grand Design

The databases we are planning to integrate are as follows. Major human gene-centred databases (RefSeq/Locuslink (Pruitt and Maglott, 2001), SWISS-PROT (Boeckmann, B. et al., 2003)) will provide structural features of the genes to be extracted with sequence analysis as well as relations to other types of objects (e.g. cells, tissues, and their activities) in the form of text-formatted comments. The gene-centred databases for model organisms (Flybase (The FlyBase Consortium, 2003) for fruit flies) will give information of conserved genes that may also have an important role in human. The disease database (OMIM (Hamosh et al., 2002)) may give a basis for molecular-based diagnosis combined with measurement data, e.g. expression patterns. The pathway database (KEGG (Ogata, H. et al., 1999)) and Gene Ontology (Ashburner, M. et al., 2000) will be used to enrich functional features of genes, and major textbooks in biomedical sciences will be used to anchor a variety of data and knowledge. In addition, we will incorporate a data set of 13,543 human gene expression profiles across 71 normal human tissues (H-invitational data set, an international gene annotation conference held in 2002/8/25-2002/9/3) for integration with those types of knowledge. One of the representational characteristics of the data is the 'tissue distribution pattern' calculated as follows: 1,994 RNA sources in the data set were classified into 10 practical tissue categories and the category-mean concentrations were computed. We will also extract relationship as follows from the MEDLINE articles on PubMed: those between diseases and their clinical manifestations and those between genes and cell/tissue types.

We introduce appropriate representations of knowledge that can be as computable as that of data, and would work in the integration of data and knowledge. The targets to which they will be applied include, but are not restricted to, the databases listed above. As the studies on diagnostic reasoning (Joseph and Patel, 1990) clarified, experts and novices differ in the way they conceive links between given information and between invoked ideas, leading to different ability in generating and eliminating alternative hypotheses. Therefore, we may well define biomedical knowledge as recognition of relations between biomedical terms. This definition will justify the adoption of the second data format, the 'object-object correlation/similarity matrix'. We may well adopt the first format, the 'feature array' to form the basis of the correlation matrix.

With those formats, various types of knowledge buried in existing databases can be represented. Some of the knowledge can be extracted with no advanced processing (e.g. sequence elements annotated in the database, and reference to other databases' objects), and others need specific processing (e.g. sequence elements yet to be annotated in the database). However, most of the rela-

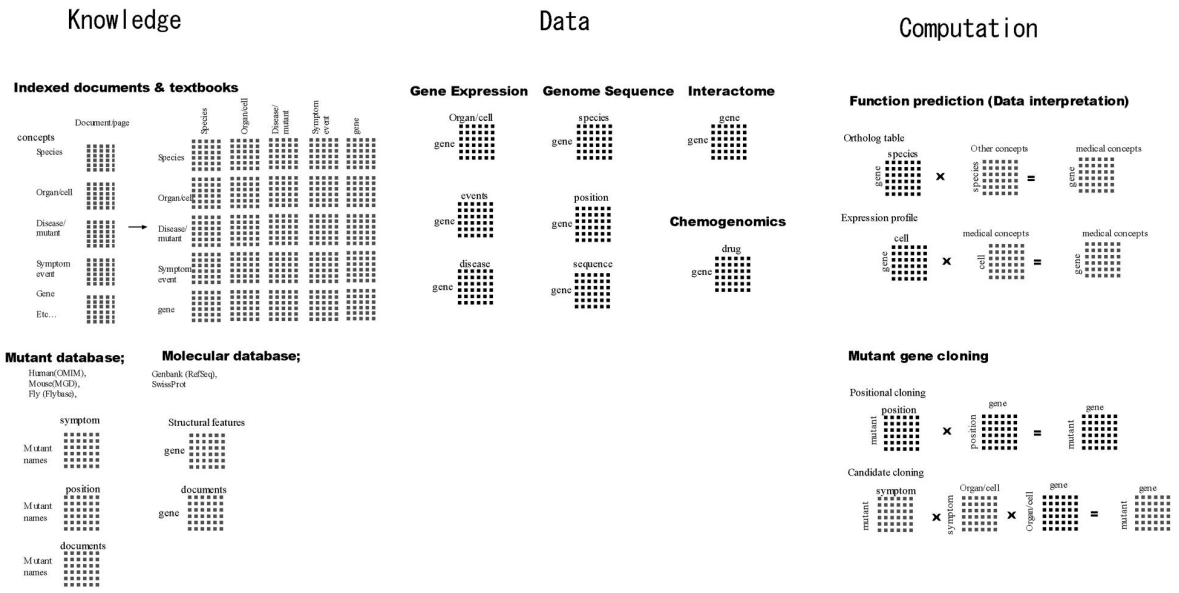

Figure 1: Representation and computation of data and knowledge. See ‘Grand Design’ section for details

tions are written in the text format with various levels of abstraction. An example of the higher-level abstraction is the statement like ‘SUBCELLULAR LOCATION Secreted’, a part of Comment section in SWISS-PROT, consisting of a feature (SUBCELLULAR LOCATION) and the value (Secreted). In this case, both the feature and the value are from a controlled vocabulary. The OMIM Clinical Synopsis (CS) field is a feature for a phenotype having a hierarchy consisting of sub-fields categorized by the type of heredity, the affected organs, or the affected systems (119 categories written in loosely controlled terms, and can be merged to about 40 categories), and under the sub-fields are about 22,000 leaf descriptions for clinical manifestations. They are composite phrase of medical terms, with about 18,000 descriptions appear only once, and the analysis of the structure and the summarization of the descriptions would be useful. An example of the lower-level abstraction is ‘FUNCTION: it induces nerve cells differentiation’, also taken seen in the Comment section of SWISS-PROT. Very basic natural language processing (NLP) would be necessary to extract the possible feature value (‘nerve cells differentiation’) for the feature (FUNCTION); moreover, the feature value may not be included in a controlled vocabulary, requiring matching of the meanings for comparison. Lastly, the lowest-level abstraction is the free text format including most of the OMIM fields and MEDLINE articles.

To cope with the textual descriptions, we will apply NLP resources as follows to the tasks: vocabularies from Unified Medical Language System (UMLS (Lindberg, 1990)) subsuming International Classification of Diseases (ICD) 10 (http://www.who.int/whosis/icd10/), a disease classification; textbook index terms; GENA (http://gena.ontology.ims.u-tokyo.ac.jp/search/servlet/gena), a vocabulary of gene names; Gene Ontology as one of the controlled vocabularies of gene functions; and GENIA (http://www-tsujii.is.u-tokyo.ac.jp/GENIA/), a biomedical and linguistic tagged corpus, to test rules to be applied to the extraction of relations between objects. These vocabularies combined with NLP techniques will extract the target objects, and co-occurrence relations between features for an object and the feature value, or between two types of objects will be calculated; then, those relations will be converted to either format: feature array or object-object correlation matrix. Once the feature arrays are calculated, they can be converted to the object-object correlation matrix. For the calculation of matrices, several sets of ontology will be necessary to match the items in rows or columns between the matrices. In addition to TissueDB, we may use

Gene Ontology to match terms representing gene functions, and UMLS to match terms representing diseases. All those processes are described in Figure 1.

3 Encoding

3.1 Disease vs Clinical Manifestations

We describe extracted relationships between diseases and their symptoms (clinical manifestations) from the MEDLINE database. In this study, we adopt a simply assumption that the frequency of the co-occurrence between disease names in titles and symptom names in abstracts indicates their strength of association.

We have conducted experiments using the whole MEDLINE database as of August 2002 containing about 12,000,000 abstracts. The dictionaries were constructed from the UMLS Metathesaurus. The disease and symptom name dictionary were constructed by gathering the terms having the semantice type of “Disease or Syndrome” and “Sign or Symptom” respectively. Since we adopt a simple longest matching algorithm for term detection, Common Engligh words such as “signs” and “other” were excluded from the dictinoaries to avoid false recognitions.

The number of unique diseases that appeared in the titles was 6,586 and that of unique symptoms was 1,083. We can thus construct a 6,586 x 1,083 matrix from the extracted pairs. Each element in the matrix represents the frequency of the co-occurrence between the corresponding disease and symptom. A disease can be represented in the form of a vector on the corresponding row.

In order to evaluate the validity of the representation of diseases by our method, we compared the similarities between diseases, which are computed as the cosine value between the vectors, with those computed by using International Classification of Disease (ICD10), which is a manually constructed disease classification. Since diseases are classified in a tree-like structure in ICD10, we define the distance of two disease on ICD10 as the number of steps along the shortest path from one disease to the other. Table 1 shows the relationship between the distance measured on ICD10 and the average similarity of vectors. The results show a clear negative correlation between them, which suggests that these two data provide similar information as for the similarity (or dissimilarity) of two diseases.

Table 1: Relationship between Distance on ICD10 and Vector Similarity

Distance on ICD10	Average of Vector Similarity
1	0.76
2	0.45
3	0.40
4	0.33
5	0.20
6	0.15
7	0.16
8	0.17
9	0.18

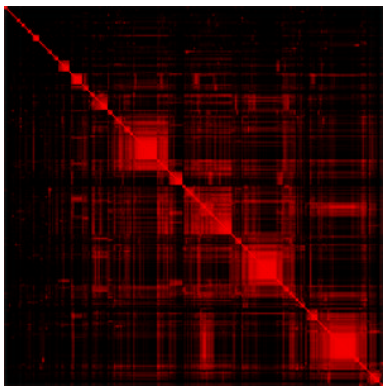

Figure 2: Result of Clustering

We next performed hierarchical clustering using the average-linkage criterion. Figure 2 shows the result in the form of a visualization of the similarity matrix. Both the rows and the columns corresond to the diseases which are sorted in the order of the clustering. The intensity of each point represents the similarity of the two corresponding diseases. The more similary they are, the brighter the point is. Therefore, the points on the diagonal are of maximal intensity because every point on the diagonal represents the similarity of identical diseases. The vague squares scattered along the diagonal indicate the existence of clusters of similar diseases.

A part of the clustering result is shown in Figure 3 in the form of a dendrogram. We can see that the diseases accompanied by seizures are merged by the clustering method.

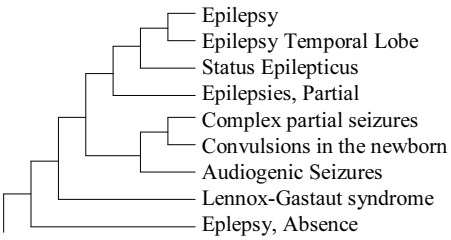

Figure 3: Part of Clustering Result

3.2 Extracting OMIM vs CS relationship from OMIM

We downloaded a text-formatted OMIM file as of August 2002, and extracted from each record the ID, title, and the Clinical Synopsis (CS) fields. Out of the 14,316 records, 4418 records had CS descriptions. We selected the subset of subfields for the CS field that represent affected body parts/organs/systems and the resultant pathophysiology (e.g. 'Oncology', standing for the malignancy caused by the mutation or accompanied by the phenotype) for 4152 records. Using the mapping information of monogenic diseases on the chromosomes provided by NCBI, we identified 990 Locuslink entries, or identified and located genes, causing 987 of the diseases. Because we found that their representation seems to be controlled only loosely (e.g. there are headings with both upper and lower case representations), we merged the headings for the selected sub-fields into 30 types.

To make a vector representation, we assumed that the pattern of the clinical manifestations of diseases could be compared in terms of how the CS categories defined here are filled. Moreover, we assumed that the comparison of each CS category was possible based on the number of descriptions in that category. In other words, a CS category may be null, filled with only one description, or filled with many descriptions, and we focused on just the tendency of descriptions distributed in the CS categories without caring for the content of the descriptions. The vectors were hierarchically clustered (Eisen, M. B. et al., 1998), and we obtained 31 tight and large clusters.

4 Application

4.1 Application of OMIM-derived disease vs clinical manifestation table

We used the H-invitational set as the source of 13,543 gene expression profiles in human adult normal tissues. They were also clustered hierarchically and 29 tight and large clusters were identified. We developed a 'cross-bar' representation that visualizes the interaction between diseases and genes (Figure 4); the rows represent diseases clustered with the patterns in affected organs/systems, while the columns represent genes clustered with the expression profiles; these patterns are represented as colored cells representing the coordinate values (for rows) or as the stacked bars representing the relative strength of expression (for columns). The intersection of a row and a column represents that the gene corresponding to the column causes the disease corresponding to the row.

One may imagine that affected organs for a disease may match with the prominent organ in the causing gene's expression pattern; however, in most of the cases, it is not true. For example, the expressions cluster #21 consisting of genes almost specific to the 'muscle/heart' tissue category causes diseases not only muscles and cardiovasucular systems, but also diseases with neurological disorders. The expression cluster #28 consisting of genes almost specific to the liver causes few diseases of the liver; rather, they cause haematological, immunological, neural, and cardiovascular diseases. It is of note that genes causing neural diseases have a wide range of expression patterns, as well as the genes causing malignancies.

4.2 Application of MEDLINE-derived disease vs clinical manifestation table

A preliminary analysis of Figure 4 showed that with the knowledge of disease cluster the prediction of the gene cluster increases by 15%. This result appears not so striking. However, we noticed in the figure that if we treat each disease cluster as one entity, the expression patterns for the causal genes are not randomly or equally distributed over the diverse ranges of expression patterns, but tend to be aggregated on some of, if not one of, the patterns; the extent of the aggregation seems to differ among

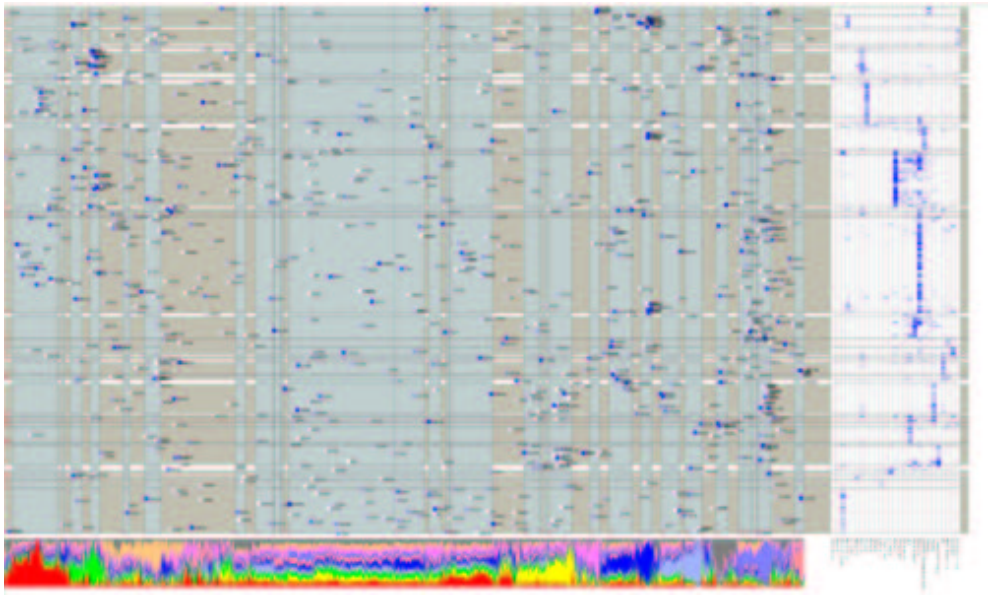

Figure 4: **Anatomical gene expression patterns and the patterns in affected organs/systems for the diseases caused by the genes.** The rows represent the diseases and the columns represent the genes. The squares placed at their intersections indicate that the gene causes the disease, and indicate the absolute abundance with the intensity of the color. Genes are clustered with their expression profiles among the tissue types, and the diseases are clustered with their patterns in affected organs/systems. The areas for tight and large clusters for either genes or diseases are colored.

the disease clusters. This implies that diseases with similar patterns in clinical manifestations may not be caused by genes with similar expression patterns, but may be caused by genes within a range of expression patterns. This property might be used to infer a set of causative genes for diseases not shown in the figure. As the first step, we are comparing the disease x clinical manifestation matrix made out of MEDLINE articles with one made from OMIM CS field in terms of their similarity in indexing.

4.3 Flybase-derived Table and Its Application

To investigate the relationship between gene function and expression pattern of the gene, we investigated association between mutant phenotype classes of *Drosophila melanogaster* (fruits fly) and expression pattern of human homologues to the fly gene.

The structure, expression and function of human genes can be studied by comparing to homologous genes of experimental animals. The homologous gene is defined as the gene of significant local sequence matching with a test sequence. The function of experimental animal genes can be determined by random or directed mutagenesis and experiments in crossbreeding. Because such information can not be acquired about human, comparison of homologues can be a powerful tool. In addition, it is well known that selecting candidate human disease genes by homology is often more successful using model organism than by considering human paralogs. In this study, we used fruits fly as a model organism, because the Drosophila has been used mutagenesis studies extensively for many decades, and genomic sequence data is available.

We made a correspondence table of Drosophila mutant phenotype with the tissue distribution patterns of the human homologues contained in H-invitational data set.

To construct this table, we converted phenotypic information of alleles in the FlyBase into a ‘Knowledge matrix’.

The conversion was carried out based on “phenotypic class:” and “phenotype manifest in:” feature in the FlyBase allele dataset. The “phenotypic class:” and the “phenotype manifest in:” feature

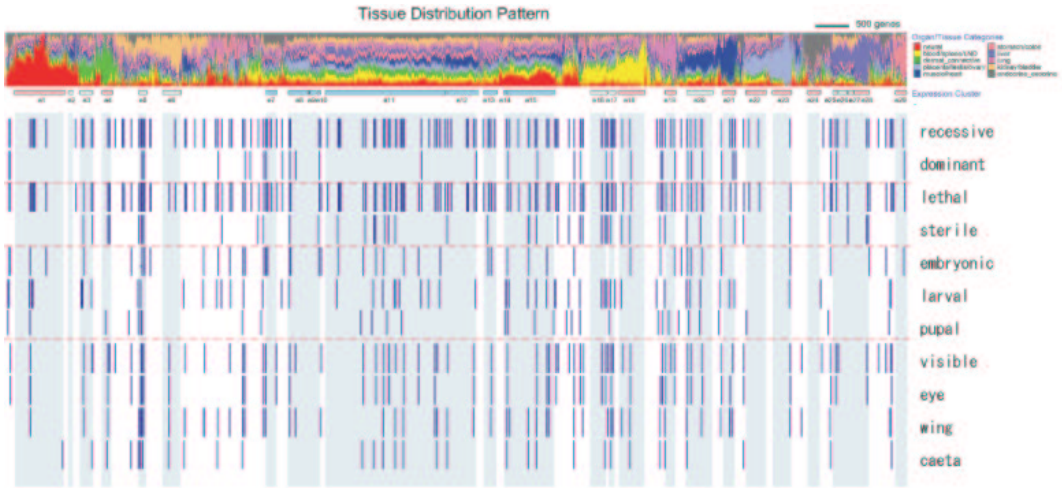

Figure 5: **Association between Drosophila phenotype classes and expression patterns of human homologues.** We used FlyBase version 3.1 dataset for our analysis. This dataset contained 42,143 alleles. 1,168 alleles had information of MIM number of human homologues. The phenotype of alleles was classified into 11 classes. The expression pattern of each gene is shown by the stacked column representing each of 10 tissue category by different colors. Numbered boxes represent 'tight' clusters e1 though e29 and colored as follows: red for 'tissue specific' and blue for 'even'. The blue bars show associations between Drosophila phenotype class and expression pattern of human homologue. The aggregation in blue bars suggests that genes for corresponding biological functions are dense in the corresponding expression clusters.

were subcategories of "phenotypic information of alleles"(*k) field.

Because, in the 42,143 FlyBase allele entry, only 1,168 entries were homologous to OMIM gene entries(2), some "phenotypic class" of FlyBase should be merged into larger classes for further analysis. However description format of "phenotypic class" and "phenotype manifest in:" was rather free and hierarchical ontology was lacked. So we re-classified the allele phenotypes into 11 major classes based on "phenotypic class:" and "phenotype manifest in:" feature (Figure 5). Each major class contained some 40 to 160 OMIM gene entries.

Human homologue of the Drosophila mutant alleles were obtained using MIM number in a "cross-reference to non-Drosophila homolog(s)/analogs"(*j) field in the FlyBase. The homologues were associated to the tissue distribution pattern (Figure 5).

Lots of Drosophila mutant strain had been constructed by random or directed mutagenesis experiments and many lethal alleles were known in Drosophila. Because wild type gene products of lethal alleles should have essential biological function, one may infer that the transcripts of such genes may distribute evenly across the tissues. But, from the Figure 5, we can see this is not true. The expression pattern of Drosophila lethal allele was widespread. That is, the expression of some lethal gene have tissue specificity, and others have not.

In the human homologues of Drosophila mutant alleles, few genes had endocrine/exocrine specific and Placenta/ovary/testis specific expression pattern, compared with others. The result was consistent with the fact that such organ was unique to vertebrates and mammals. We emphasize that our method is suitable for providing a whole view and clarifying such tendencies.

5 Future Directions

We have shown three types of seemingly qualitative data represented in the text format with different degree of structures successfully transformed into the array of object features, and have demonstrated that the matrix representation gives new insight into the global understanding of large-scale

measurement data and knowledge. We also presented a practical use of NLP techniques and possible targets to which the ontology of biological objects may be applied. We will test these resources for NLP in the process of extracting more types of relations between objects and the features.

References

Apweiler, R. et al. 2001. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29:37–40.

Ashburner, M. et al. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25:25–9.

Boeckmann, B. et al. 2003. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, 31:365–70.

H. J. Bussemaker, H. Li, and E. D. Siggia. 2001. Regulatory element detection using correlation with expression. *Nat Genet*, 27:167–71.

Camon, E. et al. 2003. The gene ontology annotation (goa) project: Implementation of go in swiss-prot, trembl, and interpro. *Genome Res*, 13:662–72.

Eisen, M. B. et al. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 1998:14863–8.

H. Ge, Z. Liu, G. M. Church, and M. Vidal. 2001. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, 29:482–6.

D. R. Gilbert, M. Schroeder, and J. van Helden. 2000. Interactive visualization and exploration of relationships between biological objects. *Trends Biotechnol*, 18:487–94.

Greenbaum, D. et al. 2001. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res*, 11:1463–8.

A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. 2002. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 30:52–55.

Ho, Y. et al. 2002. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–3.

Hsiao, L. L. et al. 2001. A compendium of gene expression in normal human tissues. *Physiol Genomics*, 7:97–104.

Ito, T. et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98:4569–74.

Iyer, V. R. et al. 99. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87.

T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–8.

G. M. Joseph and V. L. Patel. 1990. Domain knowledge and hypothesis generation in diagnostic reasoning. *Med Decis Making*, 10:31–46.

Lander, E. S. et al. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

C Lindberg. 1990. The unified medical language system (umls) of the national library of medicine. *J Am Med Rec Assoc*, 61:40–42.

E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86.

Ogata, H. et al. 1999. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27:29–34.

K. D. Pruitt and D. R. Maglott. 2001. Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Res*, 29:137–40.

A. W. Rives and T. Galitski. 2003. Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, 100:1128–33.

Scherf, U. et al. 2000. A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, 24:236–44.

A. I. Su. 2001. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A*, 99:4465–70.

The FlyBase Consortium. 2003. The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Research*, 31:172–175.

Velculescu, V. E. et al. 1999. Analysis of human transcriptomes. *Nat Genet*, 23:387–8.

Xie, H. 2002. Large-scale protein annotation through gene ontology. *Genome Res*, 12:785–94.