post-transcriptional regulators that bind to complementary sequences on target mRNAs, usually resulting in translational repression or target degradation, which then influences clinical outcome (Tseng et al., 2011).

In this chapter, we use the glioblastoma multiforme (GBM) study in TCGA as an example to illustrate our integration approach. In TCGA, copy number, methylation, mutation status, mRNA expression, and microRNA expression are measured on the same set of samples for more than 20 cancers. In the near future, RPPA data will also be available for the same patient samples. Details of TCGA projects can be found in Chapter 2, An Introduction to The Cancer Genome Atlas.

GBM is the most common and most aggressive malignant primary brain tumor in humans. For this reason, it was the first cancer type investigated in TCGA. The yearly incidence is 3–5 newly diagnosed cases per 100,000 population. Most cases of GBM develop rapidly and have a clinical history of only a few days or weeks. The overall median survival time for patients treated with the current standard chemoradiotherapy is approximately 15 months. The etiology of glioblastoma remains largely unknown, but epidemiology studies have shown that the risk factors for GBM include sex, age, and ethnicity (Preusser et al., 2011). TCGA GBM includes more than 500 GBM patient samples with DNA copy number, mutation, methylation, and gene expression information. All data sets analyzed in this chapter are publicly available and can be downloaded directly from TCGA (`cancergenome.nih.gov`). Analyzing different platforms one by one with the clinical outcome can identify the pathobiological features and molecular biomarkers in GBM. The clinical outcome we are interested in is patient survival time after diagnosis with GBM. Srinivasan et al. (2011) concluded that microRNA plays an important role in predicting GBM patient survival time.

The rest of this chapter can be summarized as follows. In Section 14.2, we explain the idea of using graphical models to integrate multiplatform data. In Section 14.3, we introduce the objective Bayesian model selection approach for Gaussian graphical models (GGMs) and describe how this Bayesian selection approach can be applied to our data set. In Section 14.4, we describe the data set of TCGA GBM and its initial preprocessing and then present detailed application results. In the end, we discuss the analysis results. In Section 14.5, we provide a summary and discussion of our results.

14.2 Graph-Based Integration of Multiplatform Data

We demonstrate our approach by analyzing the biological relationships among patient clinical outcome (overall survival time), microRNA expression, and

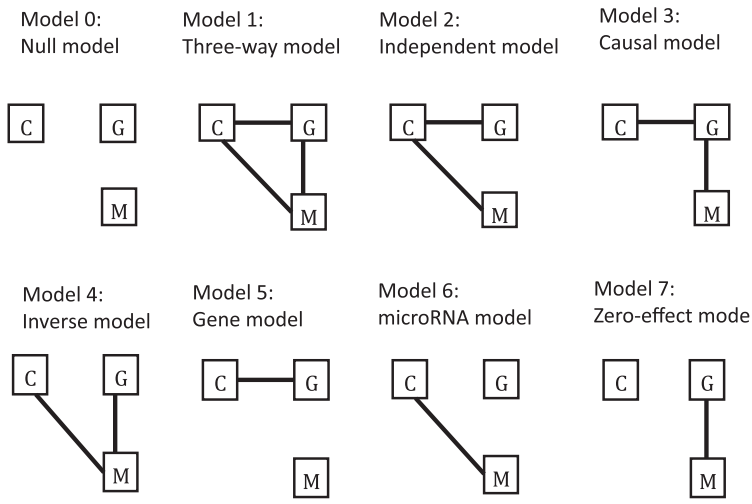

Figure 14.2 Eight possible graphical models for each G-M-C triplet (G: gene expression, M: microRNA expression, C: clinical outcome). An edge between any two nodes means that these two nodes are independent given the third node.

gene expression for GBM patients. Analyzing the relationships among other multiple platforms can be done using a similar framework, as discussed in Section 14.5. Unlike studies that integrated multiple features from different genomic alterations in one model (Zhu et al., 2004; Stingo et al., 2010; Talluri et al., 2012), we take only one feature from each platform and model the important biological relationships underlying these features and their effects on clinical outcome. Let one *triplet* denote a combination of an expression level for one gene, an expression level for one microRNA, and patient clinical outcome. There are $2^3 = 8$ possible relationships for one such triplet (see Figure 14.2). The eight models can be described as follows:

- Model 0: There is no association among microRNA expression, gene expression, and clinical outcome. We refer to this model as the *null model*.
- Model 1: Both gene expression and microRNA expression affect clinical outcome. In addition, there is an association between microRNA expression and gene expression after adjustment for clinical outcome. We refer to this model as the *three-way model*.
- Model 2: Gene expression and microRNA expression affect clinical outcome independently. There is no association between microRNA expression and gene expression after adjustment for clinical outcome. We refer to this model as the *independent model*.

- Model 3: MicroRNA expression modulates gene expression, which then affects clinical outcome. MicroRNA expression does not independently affect clinical outcome given gene expression. This relationship is consistent with the underlying biological mechanisms. We refer to this model as the *causal model*.
- Model 4: Gene expression affects microRNA expression, which then affects clinical outcome. Gene expression does not independently affect clinical outcome given microRNA expression. The relationship in this model is an inverse of the relationship in the causal model. We refer to this model as the *inverse model*.
- Model 5: Only gene expression affects clinical outcome. We refer to this model as the *gene model*.
- Model 6: Only microRNA expression affects clinical outcome. We refer to this model as the *microRNA model*.
- Model 7: Neither gene expression nor microRNA expression affects clinical outcome. The only association is between gene expression and microRNA expression. We refer to this model as the *zero-effect model*.

Of the eight possible graphical models, the three-way model, the independent model, and the causal model reflect different underlying biological processes, and biomedical researchers are particularly interested in the microRNAs and genes with such relationships. The null model, inverse model, gene model, microRNA model, and zero effect model also reflect meaningful biological processes but are relatively less interesting in our context.

Approximately 2,000 human microRNAs have been annotated to date, and this number is still increasing. Unlike the molecular features measured at the DNA level, which only modulate the mRNA expression of their corresponding genes or nearby genes, microRNAs can regulate the mRNA expression of any gene regardless of its locus, and each microRNA has multiple target genes. The relationship between a microRNA and a gene depends only on their inherent features (e.g., microRNA sequence and structure). Thus the relationships between microRNA expression and gene expression are more complicated than other regulatory networks. Currently, there are five algorithms – mirSVR, miRanda, TargetScan, PITA, and PicTar – that use the microRNA sequence and structure information to determine their target genes. However, these algorithms do not consider the effect of microRNA and gene expression on a clinical outcome.

Our goal is to determine whether the dependence pattern for each triplet follows the three-way model, independent model, or causal model. To achieve this goal, we can borrow the traditional approach to study network data, GGMs