Chapter 6

Clustering with Constraints for Integration of Heterogeneous Biological Data

The transcriptome of cells measured with microarrays gives an important and informative snapshot of the genetic information flow. However, it only reflects one particular aspect of the cell control dynamics: the number of specific RNA molecules present in a cell. Recently, several other large-scale technologies, which explore distinct aspects of the cell information flow, became available. For example, protein-protein interaction screens reveal the composition of proteins complexes [83, 108]; chromatin immunoprecipitation experiments detect where a particular protein binds in DNA genomic regions [128]; and in-situ hybridization techniques elucidate the spatial patterns of gene expression within an organism [214]. Other useful sources of large-scale data are biological databases. For example, Gene Ontology is a controlled vocabulary of biological concepts and gene annotations [9]; the Kyoto Encyclopedia of Genes (KEGG) catalogs manually annotated biological pathways [114]; and PubMed indexes titles and abstracts of most biological and medical journals [167]. Combining one (or more) biological sources of information with gene expression data is a natural next step to achieve better functional hypotheses. Indeed, several methods have been proposed for this problem (see [217] for a general review). Among others, probabilistic methods have been widely applied in this context, since they are flexible, can be easily extended to accommodate new data sources, and allow a statistical evaluation of the results [15, 192–194, 209, 218, 231].

We propose in this chapter the use of a simple, intuitive and mostly assumption-free framework of semi-supervised learning for the joint analysis of data from heterogeneous biological sources [39]. Semi-supervised learning is appropriate if there is a number of labels available for some of the observations, while the majority of data points carry no label. The main idea is to take advantage of both the labeled (supervised) and unlabeled data (unsupervised) in order to obtain better estimates than when analyzing each data source separately. For example, in [187] it was shown that few high quality labeled genes were able to improve the clustering of gene expression time courses, in comparison to a purely unsupervised method. One particular type of semi-supervised learning is called clustering

with constraints, or constrained clustering. It only makes weak assumptions about the labels by encoding secondary information as pairwise constraints. These methods search for clustering solutions, which violate the fewest number of constraints. We can, for example, derive constraints from Gene Ontology annotation (GO) [9] by constraining pairs of genes with similar GO annotation to be in the same cluster. Likewise, we can also constrain pairs of genes with distinct GO annotations to be in different clusters (negative constraints). The use of clustering with constraints for integration of heterogeneous data is based on two assumptions not explored by previous approaches [15, 192, 209, 218, 231]: (1) the secondary information is usually not available for all genes from expression experiments; and (2) gene expression data sets provide one view of the biological process under investigation, which is very unlikely to provide the same level of detail as in the secondary information. Using additional data as secondary information, we simply limit the gene expression based clustering results to biologically more plausible solutions.

In this chapter, we investigate the use of clustering with constraints for finding groups of co-expressed genes with the aid of secondary information. First, we describe related work in Section 6.1. A general formulation of the clustering with constraints problem will be introduced in Section 6.2. In Section 6.2.1 we describe the method previously proposed in [123], which we adopt in our biological applications. One contribution of this chapter is an experimental analysis of data sets commonly used in studies integrating heterogeneous biological data. The main purpose of this analysis is to evaluate the feasibility of clustering with constraints in this problem scenario [52]. We apply the clustering with constraints to yeast cell cycle data [42], using either Gene Ontology [9] or transcription factor location analysis [128] as secondary information (see Section 6.2.2 for constraints definitions). As the yeast cell cycle data set has full class labels, we can evaluate the improvements resulting from the addition of the secondary information in the analysis (see Section 6.3.1 for results). The second contribution of this chapter is a novel bioinformatics application for finding syn-expressed genes [48]. More precisely, we analyze gene expression time courses of Drosophila development using in-situ RNA hybridization images as secondary data. The constraints derived from the in-situ data are described in Section 6.2.2 and the results are presented in Section 6.3.2. Finally, we present a discussion and future work in Chapter 7.

6.1 Related Work

Semi-supervised learning (SSL) is a topic of great interest in the machine learning community [39]. SSL methods try to combine characteristics of supervised and unsupervised learning methods in problem scenarios where only part of the observations are labeled. Such data arises in many practical applications. For example, in text categorization problems, it is easy to retrieve thousands of texts from the web, but manually labeling texts is expensive [45]. Similarly, for gene expression derived from microarrays, we have the measurements of the transcription of whole genomes, but only a small fraction of genes have a

functional characterization [187]. One can implement semi-supervised learning with different machine learning paradigms [39]: transductive learning, such as transductive support vector machines [112]; graph-based approaches, such as spectral methods [207]; methods based on change of representations, which use labeled data to recompute distance matrices [118, 228]; and generative models, such as probabilistic clustering methods [19]. We are mainly interested in the latter category, as they can be used together with the mixture model framework used in other chapters of this thesis.

Semi-supervised clustering methods consider SSL from an unsupervised learning point of view. In particular, one assumes that the total number of classes and the coverage of labels in these classes are both unknown [19]. With generative models, we view the clustering problem in a probabilistic setting, and include the constraints in the model prior, in order to restrict the solution space to clustering solutions respecting the constraints derived from class labels. The semi-supervised clustering problem can be described in a complete likelihood formulation and be solved with extensions of the EM algorithm (see Section 6.2). One alternative is to use the labels as hard constraints [45, 161, 185]. A more flexible, simple and assumption free approach is to consider only constraints between pairs of objects. Most methods of clustering with constraints are based on defining a hidden random Markov field (HRMF) in the constraints [153]. They employ distinct approximation methods for estimating the posterior assignment of the EM algorithms. Among other proposals, there are: chuncklet model [196], iterated conditional modes [19], Gibbs sampling, [137], mean field approximation [123], and re-sampling chunklet model [153]. The work in [153] performed a comparative analysis of the previous methods [19, 123, 137, 153, 196] with benchmarking data sets, and with the inclusion of noise in the constraints. In general, methods like [19, 123, 153] performed well after the addition of noise, while the exact method based on hard constraints [196] had poor results. This is explained by the fact that particular sets of "hard" constraints will have no feasible solutions for a specific number K of clusters [57]. For example, the constraints in Figure 6.1 (c) cannot be satisfied for $K = 2$. Thus, exact methods should be avoided, such the one in [196], when one expects errors in the constraints. On the other hand, [153] shows that approximate methods, such as [19, 123, 137], which are based on local update rules of the posterior assignments, can get easily trapped in local maximum solutions, in particular when large constraint weights are used. The use of distinct Bayesian classifiers in a semi-supervised clustering with hard labels was proposed in [45]. The authors investigated the effects of the size of labeled and unlabeled data on UCI benchmark data sets. Their results showed that unlabeled data can deteriorate the overall results, if the assumptions of the model do not match the distribution of the data. They suggest that cross-validation on labels (or constraints) is a relevant approach for performing model selection.

Analysis of heterogeneous biological data has been tackled with several distinct methodologies. See [218] for a broad review of the area. We describe below only those studies based on semi-supervised methods. In [185, 187], it was shown how a few number of high quality labels ($< 2\%$ of observations), which were used as hard labels in a mixture model, could improve clustering of gene expression time courses. In [193], a gene expression data

set was analyzed in conjunction with protein-protein interaction data. The author also proposed a model-based clustering method with a HRMF over the protein-protein interaction graph. A belief network propagation method was used for estimation of the posteriors. In [198], pathway information from KEGG was modeled also as a HRMF, which was estimated with the interactive conditional modes method. In [194], gene expression data was analyzed together with transcription factor binding site (TFBS) data with an EM based method. Also, a model-based approach similar to [187] was proposed in [161] for clustering gene expression data with labels derived from functional annotation data. That work, however, makes an *ad hoc* selection of few functional classes used as labels, and ignores the fact that genes can be assigned to multiple functions. The same authors also investigated the use of a semi-supervised method based on the modification of the distance function according to the labeled data on similar data sets [102]. Furthermore, [189] performed a case study using the mean-field approximation for clustering with constraints [123]. They used a fully labeled yeast cell cycle data set (as in the study described in Section 6.3.1) and TFBS data for deriving the constraints. They could show that with a more conservative choice of constraints the TFBS data yielded improvements in the recovery of Gene Ontology terms.

The work presented in this chapter differs from [102, 161, 185, 187], as they are all based on hard constraints and ignore the existence of noise in the constraints. In relation to [189, 193, 198], all share a similar computational method with the one used in this thesis, but they differ in the data used as secondary information.

In the context of syn-expression, [214, 215] performed a large-scale study of gene expression in the Drosophila embryos by in-situ RNA hybridizations. The images were manually curated and annotated using a controlled vocabulary—ImaGO—following the example of the Gene Ontology [9]. The final result was a hierarchical clustering of genes based on the manual annotations; the gene expression time-courses were not included in the analysis. Recently, a similar study was performed in Drosophila embryogenesis using high-resolution fluorescent in-situ hybridization technique [127]. This technique allows the sub-cellular location of expression. They also extended the vocabulary from ImaGO to include sub-cellular location terms. Recently, studies investigated pattern formation in Drosophila based on 3D in-situ images [96, 117] for a small number of genes. Further work concentrated on mining the image database for genes with a spatial expression pattern similar to a query [160, 163] and on the extraction of relevant features in the images [160], for example by clustering images on an eigenvector based representation [162]. All these syn-expression studies restricted themselves to the analysis of the images with gene expression location. In contrast, the application proposed in this chapter is the first one combining gene expression from microarrays with gene expression location for deriving groups of syn-expressed genes.

Recently, [181] proposed the use of gene expression time courses of Drosophila development as an input for a classifier distinguishing modules of gene expression location. The modules of expression location were derived from the manual annotation of in-situ patterns

from [214] and no image processing was performed.

6.2 Mixture Model Estimation with Constraints

The main idea of clustering with constraints is to include additional data in the form of pairwise constraints in order to restrict or penalize particular cluster solutions. These constraints can be of two types: *positive* constraints, which indicate that two objects should be in the same cluster, and *negative* constraints, which indicate that two objects should be in separate clusters. Moreover, the constraints can be interpreted in two ways: "hard constraints", which have to be fulfilled in the solutions, and "soft constraints", which might be violated. For the latter, a penalty violation value can be defined for each pairs of objects. See Figure 6.1 for an example of how the "hard" and "soft" pairwise constraints can be used to restrict clustering solutions.

In this chapter, we are interested in probabilistic methods using "soft constraints" [123, 137]. One way to achieve this is to extend the basic EM approach (Section 2.3.1) to include the constraints. In the following, we describe the basic formalism of this extension. In Section 6.2.1, we describe one particular method for performing mixture model estimation with soft constraints.

Formally, for a data set $\mathbf{X}$ with N observations, we specify the positive constraints as a matrix W^+, where $w_{ij}^+ \in [0, \infty]$ is the positive constraint penalty for the pair of observations i and j ($1 \leq i \leq N$ and $1 \leq j \leq N$). Likewise, we specify a negative constraints matrix W^-, where $w_{ij}^- \in [0, \infty]$. We use W to denote the pair (W^+, W^-). Recalling Section 2.3.1, the EM algorithm is based on maximizing the complete data likelihood (Eq. 2.7),

$$\mathbf{P}(\mathbf{X}, \mathbf{Y}|\Theta) = \mathbf{P}(\mathbf{X}|\mathbf{Y}, \Theta)\mathbf{P}(\mathbf{Y}|\Theta),$$

where $\mathbf{Y}$ indicates the cluster assignments of observations in $\mathbf{X}$.

The constraints can be added into the previous equation making the prior of the cluster assignments $\mathbf{Y}$ to be dependent on W,

$$\begin{aligned} \mathbf{P}(\mathbf{X}, \mathbf{Y}|\Theta) &= \mathbf{P}(\mathbf{X}|\mathbf{Y}, \Theta)\mathbf{P}(\mathbf{Y}|\Theta, W), \\ &= \mathbf{P}(\mathbf{X}|\mathbf{Y}, \Theta)\mathbf{P}(\mathbf{Y}|\Theta)\mathbf{P}(W|\mathbf{Y}, \Theta). \end{aligned}$$

The only term depending on the constraints is $\mathbf{P}(W|\mathbf{Y}, \Theta)$. This can be interpreted as a weighting function penalizing cluster assignments $\mathbf{Y}$, which violate the constraints W. As it is common in probabilistic clustering with constraints methods [39], we assume that the constraints impose a hidden Markov random field (HMRF) on the (hidden) variable Y representing the unknown cluster assignments. In short, a hidden Markov random field is a

graphical representation of the joint distribution of a hidden variable. The HMRF assumes that the conditional distribution of the variables obeys the Markov property, i.e., the probability of a variable is only dependent on neighboring variables (see [131] for a complete description of HMRF). In our context, the HMRF graph is represented by a set of nodes, where node i represents the observation y_i, and the neighborhood graph is represented by the constraints, where w_{ij} indicates the weight of the edge between nodes i and j. Hence, it follows from [92] that the prior probability of a particular cluster assignment $\mathbf{Y}$ follows a Gibbs distributions,

$$\mathbf{P}(W|\Theta, \mathbf{Y}) = \frac{1}{Z} \exp\left(\sum_{i}^{N}\sum_{j\neq i}^{N} -w_{ij}^{+}\mathbf{1}(y_j \neq y_i) - w_{ij}^{-}\mathbf{1}(y_j = y_i)\right), \tag{6.1}$$

where $\mathbf{1}$ is the indicator function and $Z = \sum_{\mathbf{Y}\in\mathcal{Y}} \mathbf{P}(W|\Theta, \mathbf{Y})$ is the normalizing function.

In this formulation, however, we cannot assume independence between cluster assignments $\mathbf{Y}$ in the E-step, as it is required by EM algorithm (Section 2.3.1). Exact inference of the posterior would now require the complete evaluation of the following equation

$$\mathbf{P}(y_i = k|\mathbf{X}, \Theta, W) = \sum_{\mathbf{Y}\in\mathcal{Y}_{y_i=k}} \mathbf{P}(\mathbf{X}|\mathbf{Y}, \Theta)\mathbf{P}(\mathbf{Y}|\Theta, W), \tag{6.2}$$

where $\mathcal{Y}_{y_i=k}$ is the space of all cluster assignments $\mathbf{Y}$ and y_i is fixed to the value k. Several approximations have been proposed for estimating the posterior, such as the chunck-let model [196], iterated conditional modes [19], Gibbs sampling, [137], and mean field approximation [123]. We adopt the approach in [123], as it allows for modeling soft-constraints, does not require sparsity of the matrices W^+ and W^-, and performs well on benchmarking [153].

Note also that in this formulation, as $\mathbf{P}(\mathbf{X}|\mathbf{Y}, \Theta)$ is independent of W, no modification is required in the M-Step of the EM algorithm. As a result, the component models proposed in Chapter 4 and 5 can be used in this clustering with constraints setting.

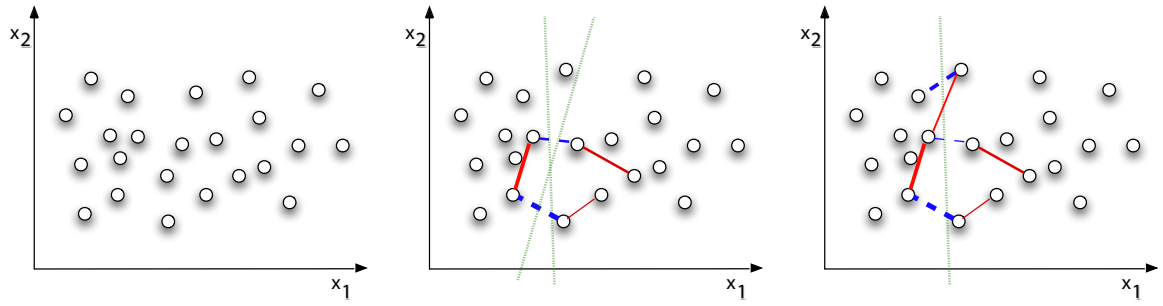
6.2.1 Mean Field Approximation

It was shown in [123] that the distribution in Eq. 6.1. follows the Maxent principle,

$$\mathbf{P}(W|\Theta, \mathbf{Y}) = \frac{1}{Z} \exp\left(\sum_{i}^{N}\sum_{j\neq i}^{N} -\lambda^{+}w_{ij}^{+}1\{y_j \neq y_i\} - \lambda^{-}w_{ij}^{-}1\{y_j = y_i\}\right)$$

where λ^+ and λ^- are Lagrange parameters defining the penalty weights of positive and negative constraint violations.

A mean field approximation is used in the inference of the posterior distributions from



(a) (b) (c)

Figure 6.1: *The effectiveness of the use of pairwise constraints, cases (b) and (c), is shown by contrasting them with the unsupervised case (a). Assuming a two-dimensional space, it is hard to distinguish the two clusters from the data points alone, and the boundary between them (a). The addition of positive pairwise constraints, depicted as red edges, and negative constraints, depicted as blue edges (b), indicate the existence of two or more clusters and possible cluster boundaries, depicted as green dotted lines. In (c), there is no boundary, which respects all constraints, and methods based on "hard" constraints would fail in this scenario. With the use of "soft" constraints, where the penalty of constraint violation is proportional to the edge widths, there is an optimal solution (green dotted line), which violates one positive constraint, in the cost of respecting a negative constraint with higher penalty value (or edge width).*

the given HMRF. Formally, the posterior distribution is approximated with a factorial distribution $q(\mathbf{Y}) = \prod_{i=1}^{N} q_i(y_i)$, by minimizing the relative entropy of the real posterior distribution $\mathbf{P}(\mathbf{Y}|\mathbf{X}, \Theta, W)$ (Eq. 6.2),

$$q^* = \underset{q}{\operatorname{argmin}} \sum_{\mathbf{Y} \in \mathcal{Y}} q(\mathbf{Y}) \log \left(\frac{q(\mathbf{Y})}{\mathbf{P}(\mathbf{Y}|\mathbf{X}, \Theta, W)} \right)$$

where $\sum_{k=1}^{K} q_i(y_i = k) = 1$.

As demonstrated in [123], the posterior assignments is approximated as follows

$$q_i(y_i = k) = \frac{\alpha_k p(x_i|y_i = k, \theta_k)}{\sum_{k'=1}^{K} q_i(y_i = k')} \exp \left(\sum_{j \neq i} -\lambda^+ w_{ij}^+ (1 - q_j(y_j = k)) - \lambda^- w_{ij}^- q_j(y_j = k) \right).$$

where α_k is defined as in Eq. 2.22 and $p(x_i|y_i = k, \theta_k)$ is the pdf of the component model (see Eq. 2.24 for the multivariate Gaussian case).

Note that this formulation allows several alternatives regarding the use of constraints. When there is no overlap in the annotations, or more precisely $w_{ij}^+ \in \{0, 1\}$, $w_{ij}^- \in \{0, 1\}$, $w_{ij}^+ w_{ij}^- = 0$, and $\lambda^+ = \lambda^- \sim \infty$, we obtain hard constraints. Alternatively, by fixing $\lambda^+ = 0$ (or $\lambda^- = 0$), we make use of only positive (or negative) constraints.

6.2.2 Deriving Constraints

We describe in this section how we can derive constraints from biological information.

Gene Ontology

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases [9]. Three structured controlled vocabularies (ontologies) describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Cellular component describes biological compartments in which genes are active (e.g., *rough endoplasmic reticulum*); molecular function contains concepts related to gene function (e.g., *catalytic activity*); and biological process describes the processes that a gene can take part of (e.g., *cellular physiological process*).

Formally, a given Gene Ontology (GO) is represented by a directed acyclic graph (DAG), in which each node t_i in a set $T = \{t_1, ..., t_M\}$ represents a biological term (controlled vocabulary or GO term) and the edges stand for relationships among these terms. A relationship $R(t_i, t_j) \in \mathcal{R}$ indicates that term t_i is a parent of term t_j. Such a relation is interpreted as t_j being a subclass of t_i, i.e., t_i is a more general concept than t_j. For instance, the biological term "*cell cycle*" is related to the more specific terms "*mitotic cell cycle*" and "*meiotic cell cycle*".

A set of genes $G = \{g_1, ..., g_N\}$ is related to a given GO term by an annotation set $\mathcal{A}$, where $A(t_i, g_n) \in \mathcal{A}$ indicates that gene g_n is annotated with term t_i. Genes often have multiple biological roles, hence they are usually annotated with several GO terms. Furthermore, the parent-child relation of GO implies that genes annotated with a term are also annotated with all parents of this term. That is, for all $R(t_i, t_j) \in \mathcal{R}$, given a gene g_n, $A(t_j, g_n)$ implies that $A(t_i, g_n)$.

The intuition for the use of Gene Ontology as a secondary data is that genes participating in the same biological process should be co-expressed [71]. Hence, we positively constrain genes annotated with the same GO terms, and negatively constrain pairs of genes annotated with distinct GO terms.

More formally, let $D(g_i) = \{t | A(t, g_i) \in \mathcal{A}, t \in T\}$ be the set of GO terms annotating g_i. We can define constraints by calculating the number of GO terms common to a pair of genes. That is, for all pairs of genes g_i and g_j (corresponding to the observations x_i and x_j in $\mathbf{X}$), we define the following constraints

$$w_{ij}^{+} = \frac{\#D(g_i) \cap D(g_j)}{\#D(g_i) \cup D(g_j)}, \tag{6.3}$$

and

$$w_{ij}^{-} = \frac{\#D(g_i) \uplus D(g_j)}{\#D(g_i) \cup D(g_j)}. \tag{6.4}$$

where w_{ij}^{+} will take values in $[0, 1]$ with $w_{ij}^{+} = 1$ indicating perfect agreement for positive constraints and $w_{ij}^{-} = 1$ perfect disagreement for negative constraints. Non-annotated genes have constraints equal to zero.

Location Analysis

Location analysis allows the detection of the binding sites of transcription factors (TF) in a genomic scale [128]. The binding of a TF to an upstream region of a gene is a pre-requisite and indicator that regulation occurs. Similarly as in the case of Gene Ontology, pairs of genes being bound by the same transcription factor are likely to be co-regulated [212].

For a set of transcription factors $F = \{f_1, ..., f_M\}$, location analysis will return relations $A'(f_l, g_i) \in \mathcal{A}'$, which indicates that factor f_l binds to g_i. Let $D(g_i) = \{f_m | A'(f_m, g_i) \in \mathcal{A}, f_m \in F\}$ be the set of TFs bound to g_i. Then, we can use Eq. 6.3 and Eq. 6.4 to obtain constraints.

In-Situ Images

An important aspect of gene expression, which has been studied in great detail in embryonic development of Drosophila melanogaster [214], is its precise localization. While the initial motivation for these sensitive experiments is to understand the role of individual genes in organ development, we can incorporate spatial expression patterns with gene expression time courses from microarrays for improving the generation of functional hypotheses.

In fact, genes that share the same temporal-spatial expression pattern are more likely to form a functional module [157]. If they are synchronously co-expressed in one tissue, or in multiple tissues, this is refereed to as *syn-expression* [157]. The spatial expression patterns can be determined with in-situ experiments where a mRNA-specific stain is produced by mRNA-binding oligonucleotides and a suitable dye [211]. Then, image analysis produces either 2D or 3D images of spatial patterns of gene expression. Drosophila embryos are morphologically rather simple, however the image analysis task is not trivial as in-situ images are taken of many subjects with large fluctuations in shape. In addition, the staining intensity has higher, gene-specific error rates compared to DNA microarrays [214].

To compare in-situ hybridization patterns of a pair of registered embryo images [159], we compute the Pearson correlation as a co-location index, as proposed in [159]. This index takes both the spatial distribution and the strength of hybridization into account. Despite its simplicity, this index had comparable performance to a more complex method previously described in [163].

More formally, let Z be an L-dimensional continuous random variable defining the pixel intensities of an image with L pixels. For a data set of images $\mathbf{Z}$, where z_i and z_j describe the pixel intensities of two registered embryo images; and z_i is an L-dimensional vector

$(z_{i1}, ...z_{il}, ..., z_{iL})$, the Pearson correlation coefficient is calculated as follows

$$\text{PC}(z_i, z_j) = \frac{Cov(z_i, z_j)}{\sqrt{Var(z_i)}\sqrt{Var(z_j)}}, \tag{6.5}$$

where $Var(z_i) = \sum_{l=1}^{L}(z_{il} - \mu_i)^2/L$, $Cov(z_i, z_j) = \sum_{l=1}^{L}(z_{il} - \mu_i)(z_{jl} - \mu_j)/L$, and $\mu_i = \sum_{l=1}^{L} z_{il}/L$.

Note that there is no annotation of the orientation of the embryo. Furthermore, automatic registration of the image is a difficult task. Hence, for each pair of images, we estimate the correlation between all possible orientations and take the maximum correlation value.

For a given gene, we have in-situ images for several developmental periods, and for each period and gene we have zero or more in-situ images. Formally, let $I_i = \{I_i^1, ..., I_i^t, ..., I_i^T\}$ indicate the sets of in-situ images related to gene i and time periods 1 to T, and let $I_i^t = \{z_1, ..., z_m, ..., z_M\}$ be the set of images related to gene i at period t. For a pair of genes and a developmental period, we compute the Pearson correlation (Eq. 6.5) for all pairs of images in sets I_i^t and I_j^t; and keep the maximum value. This yields the co-location index (CL)

$$\text{CL}(I_i^t, I_j^t) = max_{z_m \in I_i^t, z_n \in I_j^t} \text{PC}(z_m, z_n). \tag{6.6}$$

By an inspection of the distribution of the co-location index, we select a value s of gene pairs to constrain. In other words, for all pairs of genes (i, j) at period t, the sth highest $\text{CL}(I_i^t, I_j^t)$ values are positively constrained ($w_{ij}^{t+} = 1$). Similarly, the pairs (I_i^t, I_j^t) with sth lowest CL values are negatively constrained ($w_{ij}^{t-} = 1$). Using this criterion, we obtain a constraint matrix W^{t+} (or W^{t-}) for a particular developmental period t.

As a last step, we need to combine the constraints from the distinct developmental periods. See Figure 6.2 for an example. We require that a pair of genes is only constrained if it is constrained in at least p developmental periods

$$w_{ij}^{+} = \begin{cases} 1, & \sum_{t=1}^{T} w_{ij}^{t+} \geq p \\ 0, & \text{otherwise} \end{cases}, \text{and} \tag{6.7}$$

$$w_{ij}^{-} = \begin{cases} 1, & \sum_{t=1}^{T} w_{ij}^{t-} \geq p \\ 0, & \text{otherwise} \end{cases}. \tag{6.8}$$

6.3 Experiments

In this section, we describe the application of clustering with constraints in two different data sets. In the first case, for a proof of concept evaluation, we use a simple benchmarking data set—yeast during cell cycle—, which is also analyzed in Chapter 4. We use either Gene Ontology or location analysis information as secondary data. For the case of the

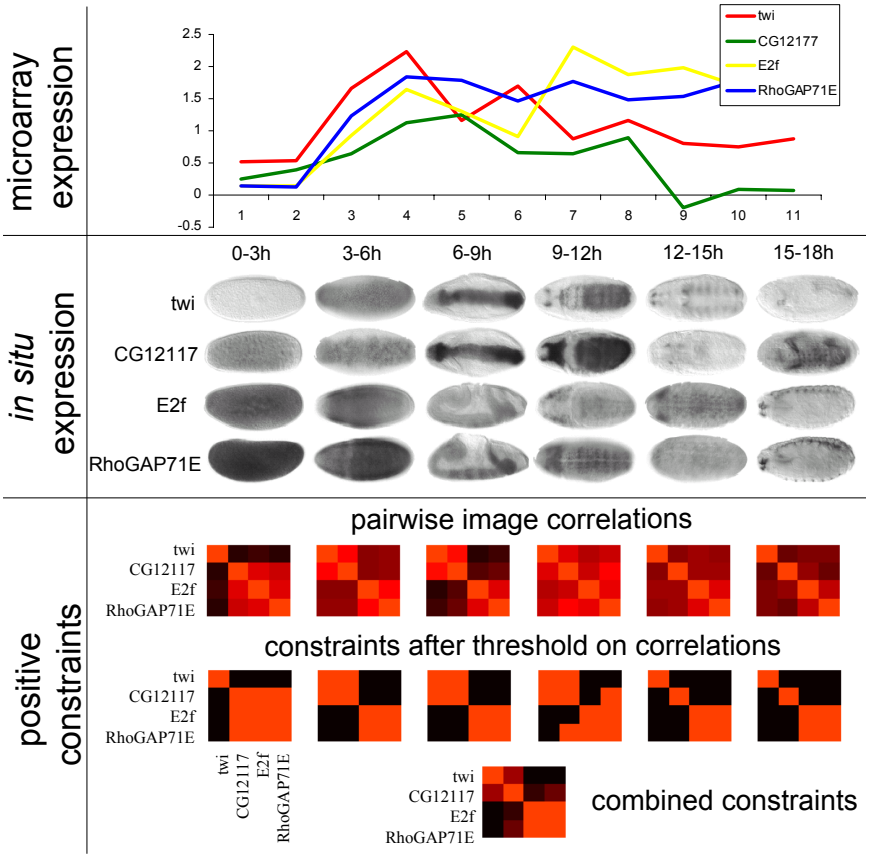


Figure 6.2: *Time course expression (top) and registered in-situ images (middle) of 4 genes twi, CG12177, Ef2 and RhoGAP71E indicate the gene expression patterns. From left to right, the embryo images are categorized into the time periods 0-3, 3-6, 6-9, 9-12, 12-15 and 15-18h. The time-courses display similar expression patterns with maximal expression after 3 hours for all genes, but weakly diverging pattern at later time points. The in-situ images indicate that twi and CG12177 have syn-expression at time periods 3-6, 6-9 and 9-12h; while Ef2 and RhoGAP71E have syn-expression at time periods 0-3, 3-6, 6-9, 9-12 and 15-18h. At the bottom, we display how positive constraints are derived from in-situ hybridization patterns. Heat-maps display the correlation coefficients between all pairs of in-situ images of the corresponding time period (red values indicate positive correlations). A constraint matrix for each time period is obtained by thresholding the corresponding correlation matrix. For example, constraint matrices from periods 3-6 and 6-9h indicate syn-expression of pairs (twi, CG1217) and (Ef2, RhoGAP71E), whereas the constraint matrix from period 9-12h indicates that (CG1217, RhoGAP71E) are syn-expressed. Matrices are combined into one, which constrains genes that display syn-expression in at least 3 periods, as indicated in the matrix at the bottom.*

second data set, we present a more detailed and exploratory analysis of Drosophila development. In this context, we use gene expression time courses as the main data set and information from in-situ images as secondary data.

6.3.1 Yeast Cell Cycle with Gene Ontology and Location Analysis

We use the expression profiles of 384 genes during Yeast mitotic cell division assigned to one of the five cell cycle phases classes [42], which we refer to as YCC. See section 4.5.1 for a detailed data description. Although this data set is biased towards profiles showing periodic behavior, and some of the class assignments are ambiguous, it is one of the few data sets with a complete expert labeling of genes.

The relation between regulators and target genes are obtained from large-scale location analysis, comprising data from 142 candidate TFs [128]. Relations $A'(f_l, g_i) \in \mathcal{A}'$ are obtained after thresholding the confidence that the TF binds to a particular gene as performed in the source literature [128]. We will refer to this data as TR.

In relation to GO, the SGD Saccharomyces cerevisiae annotation [195] is used, and for simplicity, we only included the DAG molecular process in our analysis.

Results

Multivariate normal distributions with diagonal covariance matrix are used as component models of the mixture model (see Section 2.3.3). We initialize the EM algorithm with random models, as described in Section 2.3.2. For all experiments, we vary values of λ^+ and λ^-. We use the class labels to compute sensitivity (Eq. 3.13), specificity (Eq. 3.14) and corrected Rand (Eq. 3.12).

As a proof of concept, we use the class labels from YCC to generate pairwise constraints for 5% of all pairs of genes—positive if the genes belong to the same class, negative otherwise—and observe the performance of the method with distinct constraints settings (Figure 6.3 top). In all cases, CR, Spec and Sens tend to one for λ near ten, with the exception of the experiments with positive constraints. In this case, one of the five clusters always remains empty, and two classes are joined in one single cluster. Furthermore, the use of positive constraints only has a stronger effect on the sensitivity, while the negative constraints affect the specificity. This is expected since positive (negative) constraints only penalize false negatives (false positives). It also explains the merged classes in the experiments with positive constraints, since the secondary data gives no penalty for merging two classes.

We observe similar results with GO and TR as secondary data. There is a slight but significant increase of CR and Sens for the methods with positive constraints (t-test indicates an increase at $\lambda^+ = 0.5$ with p-value $= 2.38e - 10$). However, for high λ^+ values (> 0.7), CR and Sens values decrease. No improvements are obtained with the use of positive and

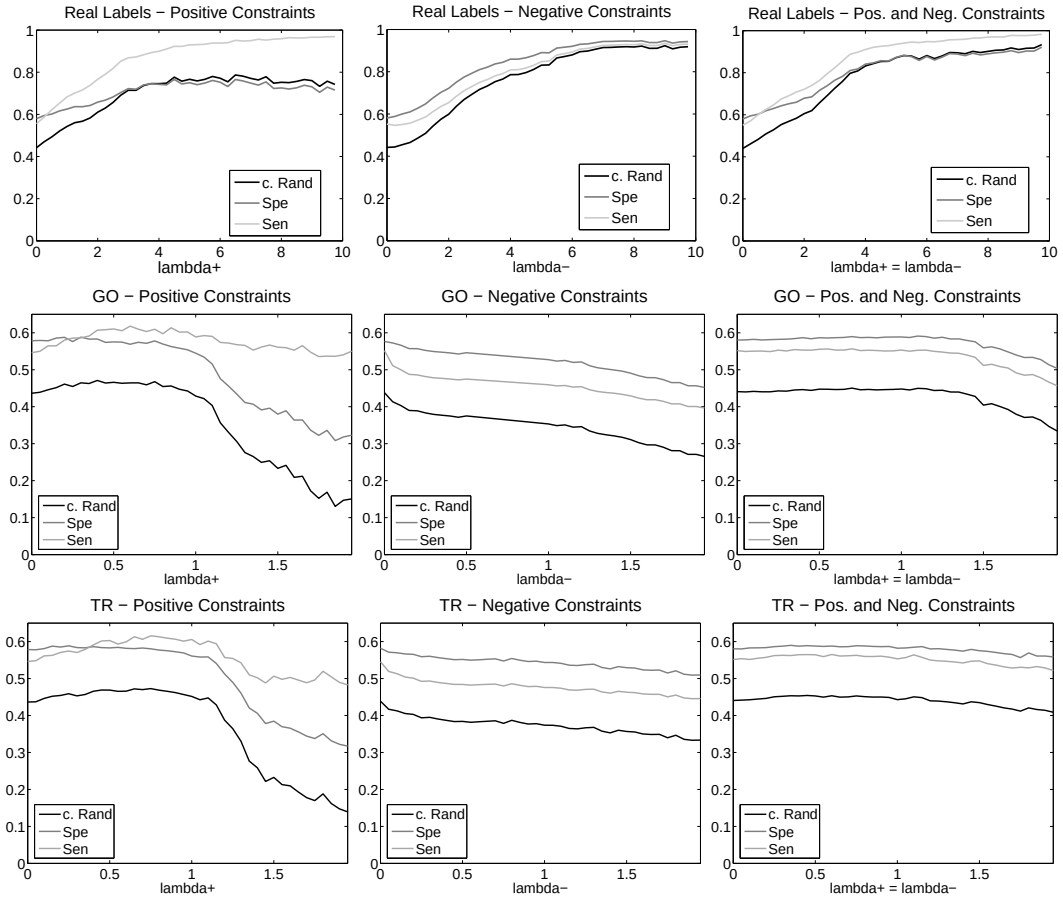


Figure 6.3: *We depict the CR, Sens and Spec after clustering YCC with positive (left), negative (middle) and positive and negative (right) constraints. We used either real class labels (top), GO (middle) or TR (bottom) as secondary information.*

negative constraints, and the negative constraints alone only deteriorate the results.

In order to understand these results, we repeat the experiments with real labels, but this time including also random labels. In total, we generate constraints for 5% of gene pairs. As seen in Figure 6.4, the addition of random labels have a great impact on the recovery of the clusters. The inclusion of 20% of random labels deteriorate the results considerably. For $\lambda = 5$, we have a CR near 0.45 for the data with 20% of noise and a CR near 0.75 for the data with no noise in the constraints. For 60% of random labels, the corrected Rand displays a behavior similar to TR and GO, obtaining low CR values (< 0.2) for high λ (> 5.0). This indicates that (1) the method is not robust with respect to noise in the data, and (2) indicates the presence of noise or non-relevant information in TR and GO.

This is not too surprising, therefore we attempt to estimate the maximal positive effect one can obtain from this secondary data. We perform the computation of enrichment analysis [24] for GO term and TR enrichment, a procedure commonly used in cluster validation, to obtain informative terms from the *true classes*. We repeat the experiments described before with the most informative TF (or GO terms) only. However, we observe only a slight improvement for the negative constraints and a relevant improvement with the use

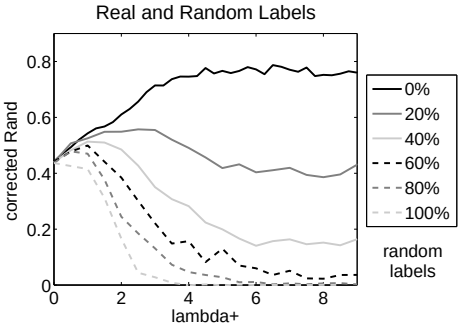


Figure 6.4: *We depict the CR obtained by clustering YCC with positive constraints from 5% of real labels with the inclusion of 0%, 20%, 40%, 60% and 100% random labels.*

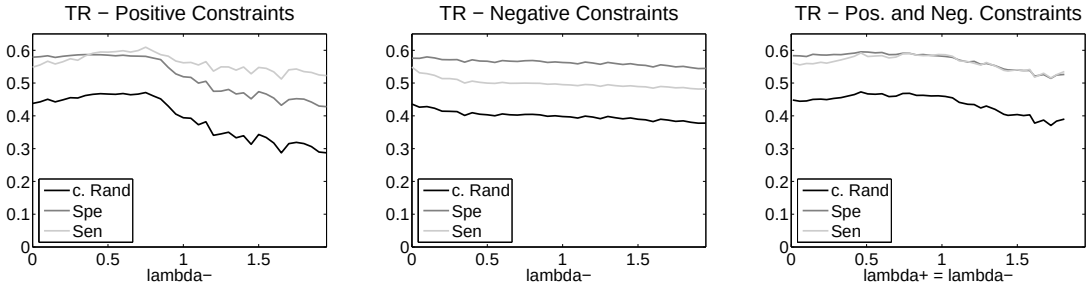


Figure 6.5: *We depict CR, Sens and Spec after clustering YCC with positive (left), negative (middle) and positive and negative (right) constraints after filtering of relevant TR.*

both positive and negative constraints in the TR data set (a CR from 0.454 to 0.472). On the other hand, no improvement is obtained after filtering terms in GO (data not shown).

These results indicate that secondary data has little power for clustering, unless it is of very high quality, free of errors and have no ambiguities. Furthermore, only as few as 20% of error in labels deteriorate the CR by more than 40%. The results for GO and TR indicate that this is the case for both biological data, and unless the procedures for obtaining constraints for GO and TR can be improved, we are more likely to deteriorate results by integrating these data. Note also that we can only obtain the best choice of λ, because the data sets are fully annotated, which is not the case of most biological data sets. Furthermore, high values of λ deteriorate results.

6.3.2 Drosophila Syn-Expression

Data

Time Courses of Drosophila Development. For twelve consecutive one-hour time windows of embryogenesis mRNA levels are measured using the Affymetrix GeneChip Drosophila Genome array. This array targets about 14,000 genes. Results were processed with the standard Affymetrix tool suite [214]. We use the median from three biological replicates. Expression values are transformed to log-ratios by using time point 1 hour as

reference. We remove genes not exhibiting at least a two-fold change, which leaves us with 2,684 genes.

In-situ Image Processing. Embryos of Drosophila Melanogaster were collected and aged to produce embryos 0-3, 3-6, 6-9, 9-12, 12-15 and 15-18 hours old [214]. The in-situ reactions were based on a cDNA library of 2,721 clones; in the end images were collected for 1,388 genes. The difference is caused either by a failure of in-situ reactions or by a lack of tissue-specific expression. Images were taken with a dissecting microscope in different focal planes and different orientations.

We use the procedure proposed in [159] for pre-processing the in-situ images. We summarize below the main steps of this image processing pipeline. The majority of in-situ hybridization images in the BDGP database contain the projection of exactly one centered embryo [22]. However, there is a noticeable portion of images with multiple touching embryos. To exploit as many data as possible, the goal of image pre-processing is to locate and extract exactly one complete embryo from each image, even for touching embryos.

To distinguish between embryo and non-embryo pixels we estimate the local variance of gray level intensities for each pixel in a 3×3 neighborhood, following [163]. It suffices to apply a fixed predefined threshold for segmentation using variance estimates because of a homogeneous background in contrast to the embryo. To eliminate erroneous embryo regions, a sequence of morphological closing and opening using a circular mask of radius four is applied [87]. Next, the largest connected component is extracted. The resulting region may be the projection of a single complete or partial embryo or the projection of a set of multiple touching embryos. To distinguish these different cases we apply a series of simple filters based on ellipticity, compactness and area of the extracted region. For regions of multiple touching embryos we introduce a procedure to separate the individuals and to extract a single complete high quality embryo. Further details are given in [159].

The final step of image processing is to register the embryos extracted to a standardized orientation and size to allow for comparison of different expression patterns. The embryo is rotated to align horizontally to the principal axis. Then, the bounding box is scaled to a standard size. Figure 6.6 shows the steps of the image processing pipeline for one example image.

We obtain constraints as described in Section 6.2.2. The 18 developmental stages of the embryo are divided into six developmental periods (0-3, 3-6, 6-9, 9-12, 12-15 and 15-18). Given the results obtained in Section 6.3.1, we would like to have only high quality constraints. Hence, we use conservative thresholds in the procedure for deriving the constraints. More specifically, we select the value s (Section 6.2.2) so that only a small percentage of gene pairs should be constrained (less than 2% of genes with in-situ images). We observe a correlation coefficient exceeding our threshold in at least three or four developmental periods, i.e., we set $p = 3$ or $p = 4$ in Eq 6.7. See Figure 6.2 for an example of how the constraints are obtained. With support of at least three periods, there are 1,756 positive constraints within 170 genes and 2,544 negative constraints within 360 genes. With

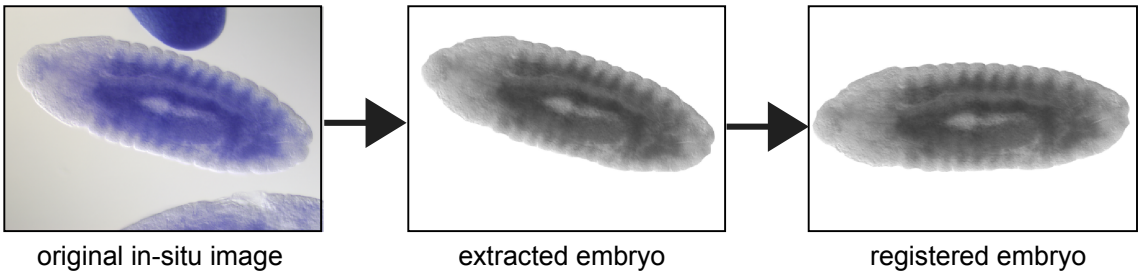


Figure 6.6: *The image pipeline combines registration, morphological operations and further processing steps to automatically process raw images, even if they include multiple touching embryos. Shown here is the image* `in-situ8784` *from gene CG5353. Image reproduced from [159]*

support of at least four stages, there are 270 positive constraints within 66 genes and 640 negative constraints within 151 genes.

ImaGO Term Enrichment. A controlled vocabulary, which follows the Gene Ontology standard [9], is used to annotate spatial gene expression patterns [214]. All images deposited in BDGP are annotated with at least one of these terms. Like with Gene Ontology enrichment analysis described in Appendix A, we can use a statistical test to list ImaGO terms that are over-represented in a cluster. Lower p-values indicate an enrichment in ImaGO terms and, consequently, better results.

This strategy is useful for evaluating the biological quality of a single cluster, but it gives no global assessment for comparing the results obtained by two clustering solutions. A heuristic to perform such an analysis is to compare the p-values obtained for two solutions [73]. A method is said to be better than another method if it has a larger number of ImaGO terms with lower p-values.

Results

We use multivariate Gaussians with diagonal covariance matrices [145] as our components in all mixture estimations. We refer to the results of the unsupervised method as `MoG` and to the clustering with constraints method as `cMoG`. We initialize the EM algorithm with random models, as described in Section 2.3.2. In the unsupervised setting, we estimate the optimal number of clusters with the BIC (Section 2.3.5), which indicates 28 clusters. We use this number for all other runs described below.

Clustering of Gene Expression Data using Mixture of Multivariate Gaussians (`MoG`). The gene expression time-courses cover the period from 1 to 12 hours of the embryo development and expression values are given as log-ratios. Overall, our clustering results reflect two typical classes (see Figure 6.7): the maternal and zygotic genes [68]. Maternal genes

appear strongly expressed in the first three hours, usually followed by a decline. Clusters 18 to 28 clearly follow this pattern. These transcripts are deposited in the oocyte; typically the embryo does not transcribe these genes in early development. They are responsible for the determination of body axes and the first phases of the cell cycle and other functions. The period from 2 to 3 hours coincides with the cellularization and the formation of three germ layers following gastrulation, when primary tissues start to develop [130].

On the other hand, genes actively transcribed in the embryo are not expressed in the early time points and expression rises to significant levels only in later stages (3 hours and later). Many of these genes are important to organogenesis. Transcripts in clusters 1 to 4, and 8 to 11 follow the pattern of embryonic activation unambiguously. The functional association can be observed in the over-represented GO terms For other clusters shapes cannot be matched to the maternal or the zygotic expression patterns. Several cluster have maximal expression in the midst of embryonic development. Note that those clusters are less populated than the ones in the maternal and in the zygotic classes.

Using in-situ Images as Secondary Information. We use semi-supervised learning to obtain better solutions for the maximum-likelihood estimation. In order to do so, we restrict the mixture estimation with constraints between pairs of genes. The principle underlying this is shown in Figure 6.1. These constraints will, ideally, differentiate between genes showing co-expression only by chance from those temporal co-expression supported by spatial co-expression (syn-expression).

We use the ImaGO enrichment analysis (Section 6.3.2) to select the best parameterization for cMoG. More precisely, we evaluate the use of constraints shared by either three or four developmental periods, the use of positive constraints and both positive and negative constraints, and four choices of the parameter λ^+ (and λ^-) (0.5, 1.0, 1.5 and 2.0) with $\lambda^+ = \lambda^-$. There is no theory guiding the choices of λ^+ and λ^-, neither is there a definitive "gold standard" or class labels to optimize them. Hence, we made the simple choice to give positive and negative constraints equal weights.

As shown in Table 6.1, all constraint combinations lead to an increase in ImaGO term enrichment, except the use of positive and negative constraints from three stages. Furthermore, values of λ around 1 lead to an improvement, while higher values tend to deteriorate the results. Thus, we choose to use the cMoG results with only positive constraints derived from three developmental periods and a constraint weight of $\lambda_+ = 1.0$.

Changes in the Biological Annotations with cMoG. To investigate the effects of the constraints in the clustering, we compare the results of MoG with cMoG (see Figure 6.8 for cMoG clusters). As explained in the previous section, we choose to use positive constraints, which are supported in at least three developmental stages, as they yield a good recall of in-situ image annotations.

As a sanity check, we inspect the number of constraints satisfied in the final solutions.

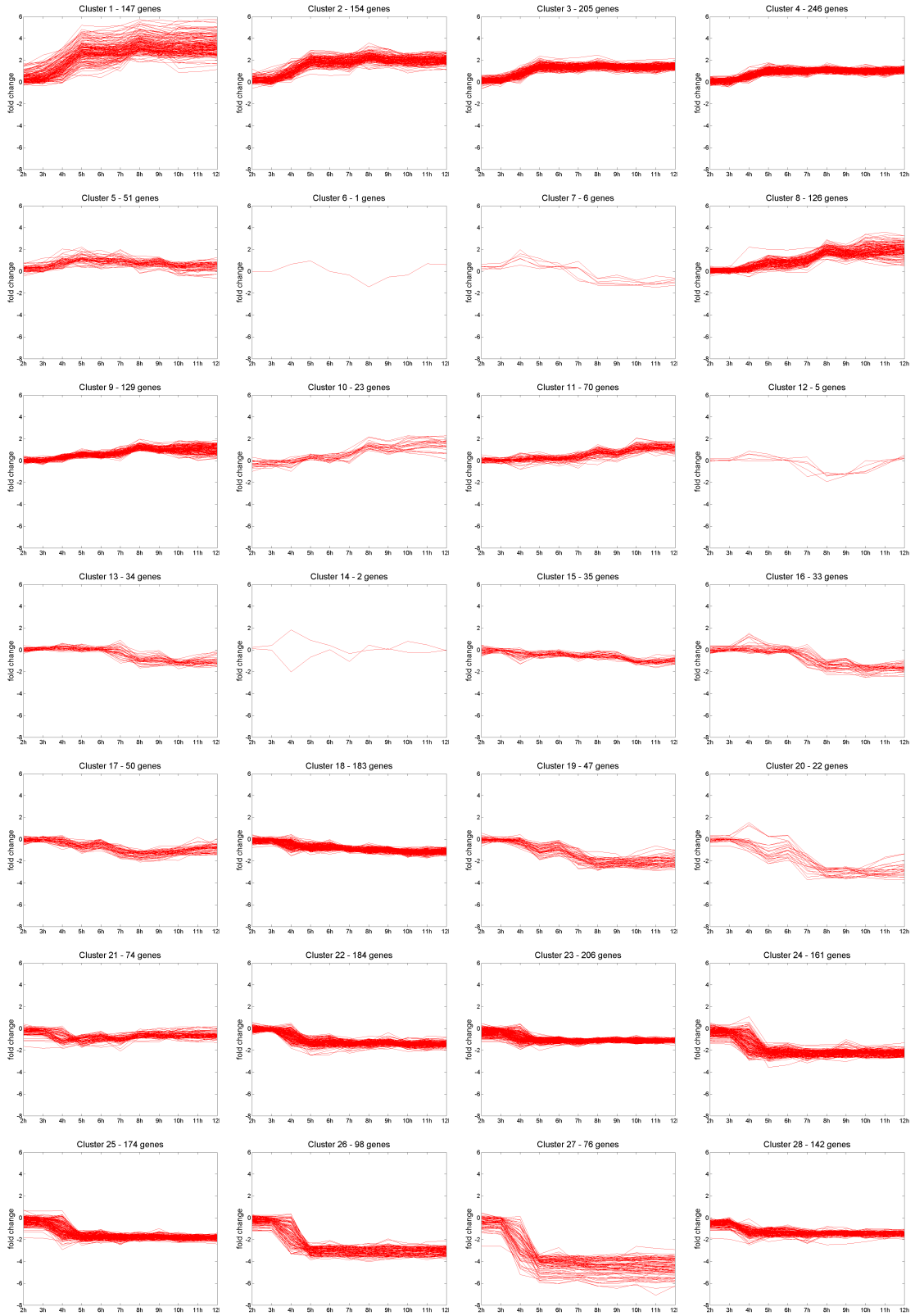


Figure 6.7: *We display the 28 clusters from* MoG.

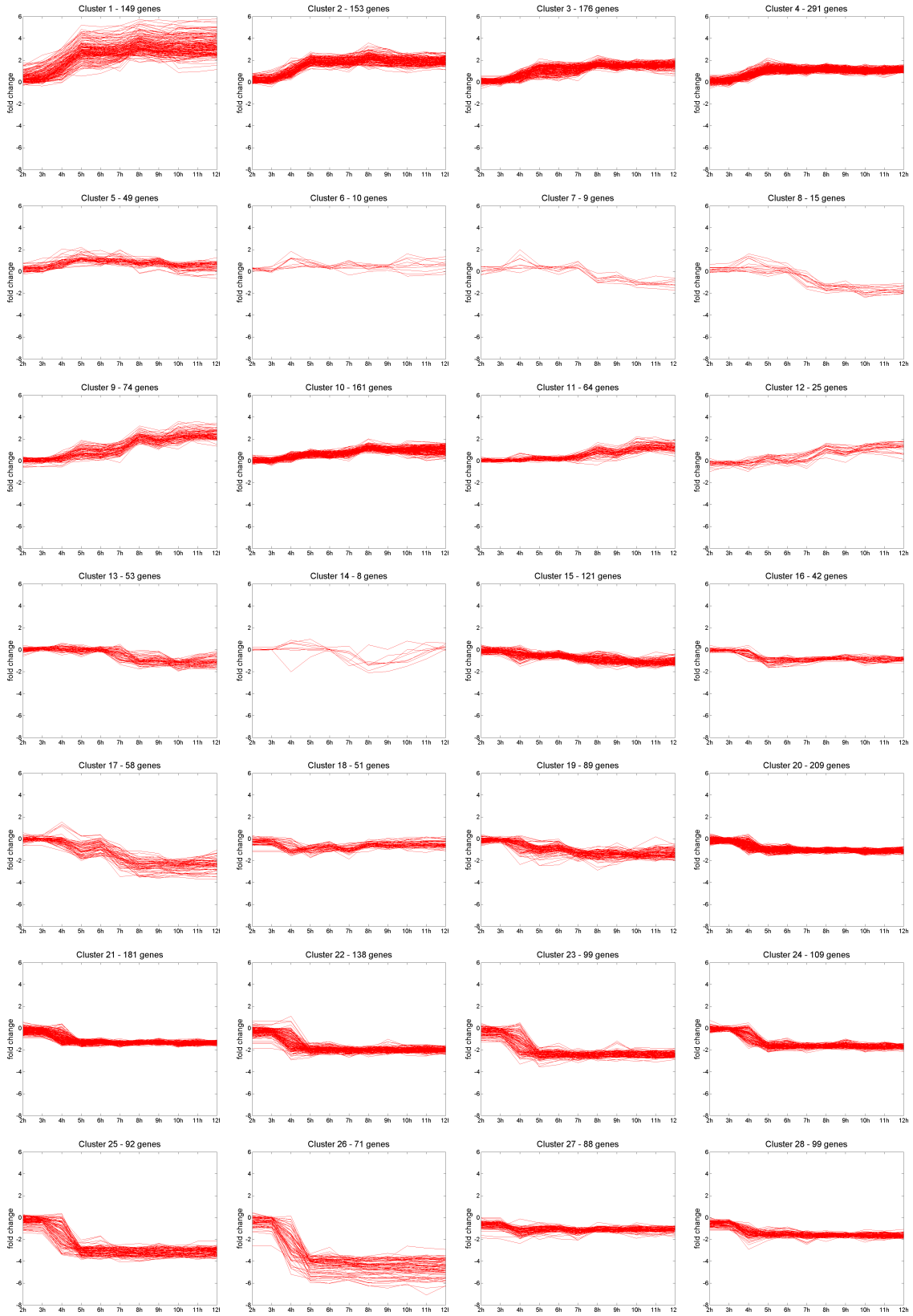


Figure 6.8: *The 28 clusters from* cMoG *show tightly co-regulated pattern and a refinement of the clustering solution of* MoG.

Table 6.1: *We compare the performance of distinct constraints and parameter choices with the ImaGO enrichment analysis. More specifically, we show the proportion of ImaGO terms with lower p-values in* cMoG *compared to* MoG *for constraints derived from a 3 or 4 stages, and distinct weights λ^+ and λ^-. Values exceeding 50% indicate an advantage of* cMoG

		Proportion of terms with lower p-values	
λ^+	λ^-	# stages ≥ 3	≥ 4
0.5	0.0	51%	48%
1.0	0.0	**60**%	**56**%
1.5	0.0	57%	49%
2.0	0.0	43%	46%
0.5	0.5	49%	44%
1.0	1.0	**49**%	52%
1.5	1.5	40%	**59**%
2.0	2.0	43%	47%

With MoG, a sizable proportion of the constraints are already satisfied (656 out of 1,756 pairwise positive constraints), as part of the expression data agrees with the constraints. With cMoG, 1,127 out of 1,756 pairwise positive constraints are satisfied. This value is nearly twice the number found with MoG. This demonstrates that cMoG benefits from the constraints in deriving the clusters of genes exhibiting `syn-expression`.

Another helpful analysis is the comparison of enrichment of in-situ image annotations (ImaGO), as described in Section 6.3.2. We display in Figure 6.9 a scatter plot of all ImaGO terms, which has an enrichment with a p-value lower than 0.01 in at least one cluster from cMoG or MoG. Based on Figure 6.9, we observe that cMoG has a higher enrichment than MoG in 67 out of 112 relevant ImaGO terms. A binomial test for testing the event of having 67 successes in 112 trials is rejected with a p-value of 0.0232, which indicates that the counts of ImaGO terms with higher enrichment for cMoG is significantly higher than expected by chance. Furthermore, if we take only ImaGO terms with a higher enrichment gain for one of the methods into account (points distant from the diagonal line in Figure 6.9), the advantage of cMoG is even greater (see Figure 6.10 and Figure 6.11). This indicates that even without direct use of the annotation information from ImaGO, cMoG has a greater sensitivity in grouping syn-expressed genes.

Overall, the individual clusters of MoG and cMoG differ only partially. Mainly, cMoG has fewer clusters a smaller amount of genes. One way to quantify the distinctions is to calculate the sensitivity and specificity of cMoG taking the results from MoG as the ground truth. These values are respectively 0.53 and 0.97, which indicate that cMoG has a tendency to subdivide clusters from MoG.

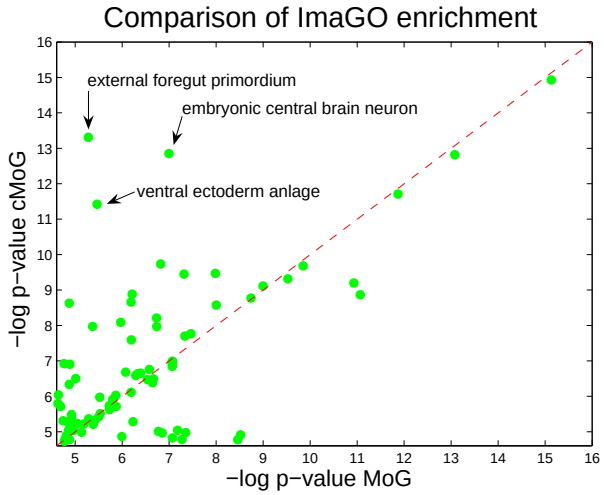


Figure 6.9: *We compare ImaGO term enrichment of* MoG *(x-axis) and* cMoG *(y-axis) in a scatter plot. We use $-log(p)$-values, thus larger values indicate a larger degree of enrichment. Points above the red line indicate a higher enrichment in* cMoG *clusters, and points below in* MoG *clusters. The distance from the diagonal is proportional to the increase in enrichment. For 67 out of 112 ImaGO terms we observe a higher degree of enrichment in* cMoG *clusters.*

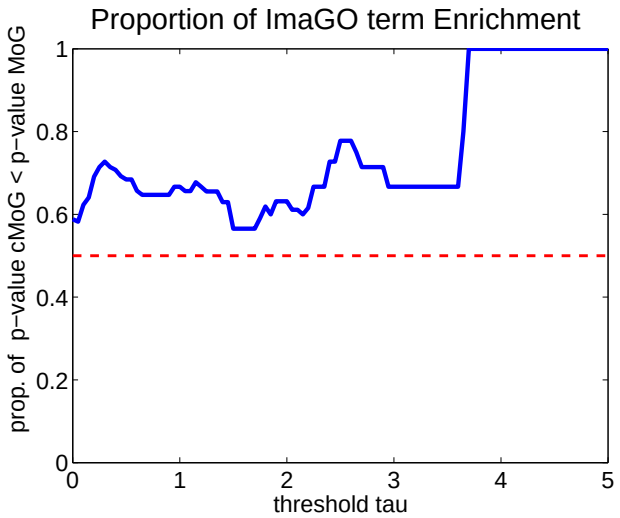


Figure 6.10: *For each threshold τ (x-axis), we depict the proportion of ImaGO terms for which we observe a smaller p-value in* cMoG *than in* MoG *(y-axis). The threshold τ discards ImaGO terms, where the difference in the log of the p-value of cMoG and MoG in smaller than τ. As can be observed, the proportions are higher than 0.5 for all τ values, which indicate an advantage of* cMoG*. Furthermore, the proportions have an increasing tendency for higher τ values.*

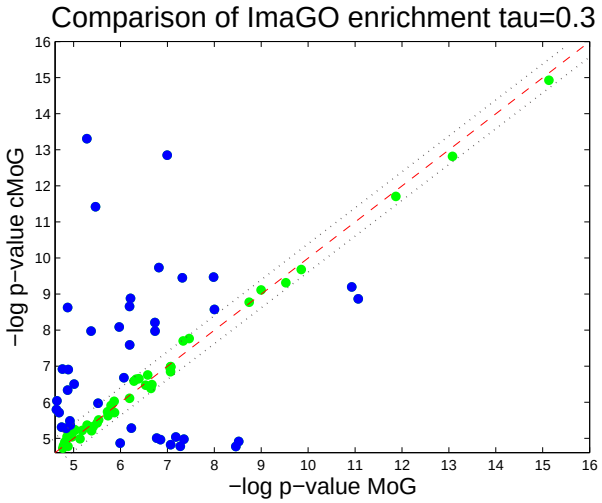
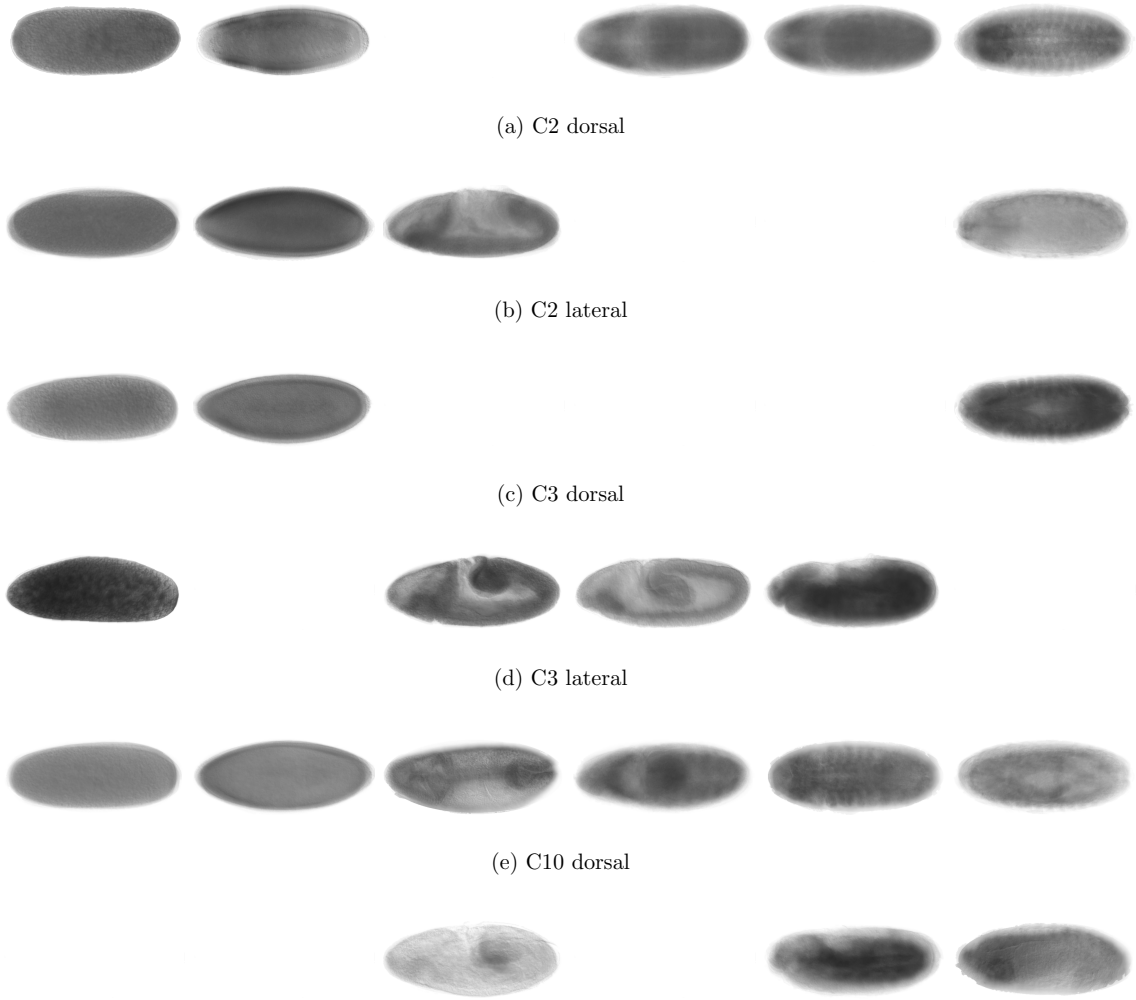


Figure 6.11: *We compare ImaGO term enrichment of* MoG *(x-axis) and* cMoG *(y-axis) in a scatter plot for $\tau = 0.3$. We use $-log(p)$-values, thus larger values indicate a larger degree of enrichment. Points above the red line indicate a higher enrichment in* cMoG *clusters, and values below in* MoG *clusters. Green points between the dotted lines represent ImaGO terms not satisfying the threshold $\tau = 0.3$, where τ indicates the distance from the diagonal line to the dotted lines. We clearly observe a higher proportion of non-filtered ImaGO terms (points in blue) above the diagonal line (32 ImaGO terms) against (12 ImaGO terms) below the diagonal. A binomial test is rejected with a p-value of 0.0018, which indicates an significant advantage of* cMoG.



(a) C2 dorsal

(b) C2 lateral

(c) C3 dorsal

(d) C3 lateral

(e) C10 dorsal

(f) C10 lateral

Figure 6.12: *Averaged in-situ images of clusters C2, C3 and C10 from lateral and dorsal views.*

Functional Annotations in `cMoG`**.** Even for a well characterized genome like Drosophila, the high dimensionality in the annotation data provides only limited information for any single gene. For evaluating the results, we need to identify the corresponding functional modules in the unconstrained and the constrained sets. It is also necessary to show improvements rather than simple correct functional assignments in either solution. In the following, we will refer to the ith cluster from `cMoG` and `MoG` as Ci and Ui respectively.

For some cases, the mapping from clusters of `cMoG` to `MoG` is simply one to one (e.g., C1 to U1, C5 to U5, C11 to U11 and C12 to U10). However, the majority of clusters show larger differences. For simplicity, we focus the functional analysis on clusters with zygoticly expressed genes (i.e., C1 to C4 and C9 to C12 in Figure 6.8).

Cluster `C2` represents a good example of the changes resulting from the introduction of

constraints. It contains most of the genes from U2 (135 genes) and 16 genes from U3. Out of the seven genes, which show similar expression patterns and have co-location constraints (*CG6930, E2f, Iswi, neur, Set, RhoGAP771e, trx*), only four (*G6930, E2f, Iswi, trx*) are found in U2. All these genes have ImaGO annotations related to *ventral nerve cord primordium* and related terms (see Figure 6.12 (a) and (b) for mean in-situ images of these genes). Related genes that have no constraints but are annotated as part of the *embryonic central nervous system* are included in C2 (*CG7372, CG14722, fzy*). The analysis of GO term enrichment returns terms such as *nervous system development* (p-value of 3.38e-23) and *system development* (p-value of 9.54e-21) (similar term enrichment is found for cluster U2). It should be noted that clusters U2 and U3 have a similar mean expression pattern. They mainly differ in the time when genes reach the plateau of maximal expression.

An example for larger changes is cluster C3, which is mainly composed of genes originally found in U3 (101 genes) and U8 (63 genes). C3 has constraints between three genes (*rhea*, *Rsf1* and *vig*) of which *rhea* and *vig* come from cluster U8 and *Rsf1* from U3 (see Figures 6.12 (c) and (d) for mean in-situ images of C3). This cluster presents higher enrichment for ImaGO terms related to *muscle primordium* (genes *CG5522, CG9253, Dg, Mef2, betaTub60D, htl, mbc, vig*) than U3 and U8. Furthermore, GO term analysis reveals that this cluster shows enrichment for *nervous system development* (p-value of 1.33e-11) and *axis specification* (p-value of 9.31e-05). For the latter term, seven genes are originally from U3 (*Dfd, Lis-1, sti, Syx1A, sqd, Ras85Dm, tup*) and five from U8 (*baz, Dg, pnt, Rac2, tok*), demonstrating that the changes introduced increase the number of syn-expressed genes within C3.

The cluster C9 represents only a subset of U8 (59 out of the 126 genes) but has no genes with constraints. It consists of genes from U8 that are not constrained to genes from C3 (see previous paragraph). Still, it is enriched in the ImaGO term *embryonic central nervous system* and related terms (genes *HLHmbeta, NetB, Oli, lin-28, scrt, sd, tap, uzip* and *zfh2*). The cluster is also enriched in the terms *organ* (p-value 2.66e-05) and *ectoderm development* (p-values 8.54e-05), which are significantly enriched in U8. In other words, this cluster is a specialization of U8, whose genes are specific to *organ development*.

C10 is formed by the addition of most genes in the U4 cluster (39 genes) to U10 (118 genes). There are seven genes constraining this cluster (*CG6751, CG18446, CG13912, CG10924, CG8745, dm, Klp61F*) (see Figures 6.12 (e) and (f)). ImaGO term enrichment relates this cluster to *yolk nuclei* and *amnioserosa*. It is also enriched in the GO term *nervous system development* (p-value 1.06e-08), all of which are insignificantly enriched in the U10 cluster.

It is also worthwhile to look at those few cases where MoG performs better. From Figure 6.9, two ImaGO terms with higher enrichment increase in MoG are *maternal* and *procephalic ectoderm anlage in statu nascendi*. The former term is enriched in cluster C22 and U21, where MoG has some more genes related to the term *maternal* (34 genes in MoG compared to 31 genes in cMoG). For the latter term, clusters U2 and C2 are both enriched, and there was only one annotated gene in U2 not in C2. As none of these annotated groups

of genes has pairwise constraints, we cannot detect any direct effect of the clustering with constraints on these results.

In summary, the refined clusters improve the generation of testable hypotheses for the role of uncharacterized genes. Overall, we observe improvement in annotation of genes related to development of the Drosophila, in particular with respect to the ImaGO annotations, which increases our confidence in the delineation of syn-expressed functional modules.