Automatic Recognition of Cells (ARC) for 3D Images of C. elegans

Fuhui Long[1*], Hanchuan Peng[1], Xiao Liu[2], Stuart Kim[2], and Gene Myers[1]

[1] Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia, USA
[2] Department of Developmental Biology, Stanford University, Stanford, California, USA
[*] longf@janelia.hhmi.org

Abstract. The development of high-resolution microscopy makes possible the high-throughput screening of cellular information, such as gene expression at single cell resolution. One of the critical enabling techniques yet to be developed is the automatic recognition or annotation of specific cells in a 3D image stack. In this paper, we present a novel graph-based algorithm, ARC, that determines cell identities in a 3D confocal image of *C. elegans* based on their highly stereotyped arrangement. This is an essential step in our work on gene expression analysis of *C. elegans* at the resolution of single cells. Our ARC method integrates both the absolute and relative spatial locations of cells in a *C. elegans* body. It uses a marker-guided, spatially-constrained, two-stage bipartite matching to find the optimal match between cells in a subject image and cells in 15 template images that have been manually annotated and vetted. We applied ARC to the recognition of cells in 3D confocal images of the first larval stage (L1) of *C. elegans* hermaphrodites, and achieved an average accuracy of 94.91%.

1. Introduction

Automatic recognition of the identities of individual cells in 3D microscopy images is indispensable for the high-throughput analysis of cellular information, such as gene expression levels and cell morphology, at the single cell level. One example is our recent work on high-throughput whole-animal single-cell gene expression analysis for *C. elegans* [1] based on a 3D digital atlas of the nuclei of this animal [2]. Currently cell recognition is accomplished by expert manual annotation, which is extremely labor intensive and basically untenable for a large number of images. Using a small set of, say a dozen or so, manually annotated images of the same organism as *templates*, we demonstrate that it is possible to extract cellular information such as location and relative spatial relationship of individual cells from these templates, and automatically assign names to cells in any new image of the same organism provided it is sufficiently stereotyped, which *C. elegans* most certainly is. But this is not the only application, for example, the embryonic and larval neurons of *D. melanogaster* are highly stereotyped, as are many other early developmental patterns. Figure 1 illustrates this problem schematically.

It is challenging to develop such automatic cell recognition technique for several reasons. First, individual cells in an image need to be segmented to high accuracy as a precursor, and this is difficult when the image quality is limited and cells are tightly clustered. Second, it is common that an image of an entire organism (e.g. *C. elegans*), or a particular tissue of an organism (e.g. the mushroom body of a fly brain) may contain hundreds or thousands of cells. Thus the scale of the problem presents a challenge to traditional graph matching techniques [4~14], which have been

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30-April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128-139, 2008.

successfully applied in applications such as face recognition [11], object tracking [12], image retrieval [13], and image registration [14] to find correspondences between two sets of spatial points, each usually containing less than a hundred objects. Finally, due to the imperfection of staining and the resolution-limit of the imaging, an expert annotator can only annotate the subset of cells in a template image that are large enough and strongly stereotyped in location. Thus the problem becomes a subset-matching problem, which is more difficult than the case where both the subject image and the template images have the same number of cells.

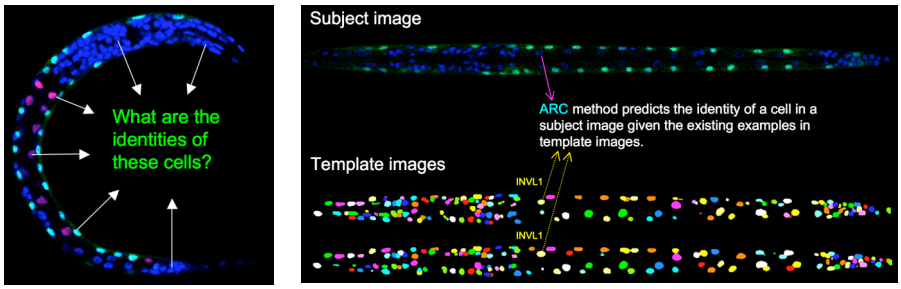

Fig. 1. (a) is a raw image of *C. elegans*, and in (b) we illustrate the cell recognition problem on image stacks where the worm has first been straightened [3], and then size and orientation normalized and segmented as described in our earlier work [2]. Cells in a template are colored so that locally it is clear which cells have the same identity between instances.

For this problem, our new method, called Automatic Recognition of Cells (ARC), is developed below in three layers of increasing refinement or power as follows. In Section 2 we introduce a basic framework of a two-stage bipartite matching that first matches cells in a subject image against the annotated cells in *each* template image, and then matches cells in the subject image to a unique cell by considering assignment scores based on the first-level matching results. In Section 3 we introduce and constrain the possible matchings to observe relative spatial invariant relationships discovered in the training stacks, specifically, the anterior-posterior (AP), left-right (LR), dorsal-ventral (DV) invariant relationships. In Section 4 we introduce a marker-based strategy in which a fiducial framework of alternately-labeled marker cells is automatically annotated with very high confidence, and then these are used to triangulate and constrain the annotation of the remaining cells.

We applied ARC to the 3D confocal images of the first larval stage (L1) of *C. elegans* hermaphrodite that has ~558 cells [1]. For this problem we manually annotated 351 cells in 15 templates. Most of the un-annotated cells are small neurons in the head of the organism. Our results show that using our marker-guided, AP/LR/DV-constrained, two-stage bipartite matching, we achieved 94.91% accuracy in searching for these 351 cells in an initially unsegmented image stack.

2. Two-stage Bipartite Graph Matching

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

Given a subject image S in which cells are to be recognized by a computer program and a template image T in which cell identities have been annotated by biologists, we can formulate our problem as bipartite matching. Consider the directed bigraph $G = (V^S \cup V^T, E)$ consisting of two disjoint vertex sets, V^S for the subject S and V^T for the template T, and all edges $E = V^S \times V^T$ from V^S to V^T. Later we will restrict E to be a subset of all the possible pairings. In the first stage we find a minimal cost, maximal matching M between V^S and V^T. That is, we minimize a cost function

$$E^1 = \sum_{s \to t \in M} D^1(s \to t) \tag{1}$$

over all sets of edges M for which adding another edge to M gives a set of edges which is no longer a matching, i.e. $\forall s \to t \left(out(s) = 1 \text{ or } in(t) = 1 \right)$ where *out* and *in* are the out- and in-degree of a vertex. $D^1(s \to t)$ is the distance between cells s and t, i.e.,

$$D^1(s \to t) = \left\| p_s - p_t \right\| = \sqrt{(x_s - x_t)^2 + (y_s - y_t)^2 + (z_s - z_t)^2} \tag{2}$$

where $p_c = (x_c, y_c, z_c)$ is the coordinate of the cell c. We find M by using the Hungarian algorithm [15].

If we have K template images T_1 through T_K, we obtain K maximal matchings M_1 through M_K against the subject S. Thus a cell in S has anywhere from 0 to K assignments of cell names. Let the set of cell labels $L = \bigcup_k V^{T_k}$ be the set of all cell names used in some template. Note carefully, that not every cell annotation in L is necessarily labeled in a template. We then use the second stage bipartite matching to determine the unique identity of each cell in S, by finding the minimum cost E^2, maximal match $M^* \subseteq V^S \times L$ with respect to the cost function D^2 defined as follows:

$$D^2(s \to t) = \left(\sum_{s \in V^S} N(s \Leftrightarrow t) \right) - N(s \Leftrightarrow t) \tag{3}$$

where $N(s \Leftrightarrow t) = \left| \{ k : s \to t \in M_k \} \right|$ is the number of times that s is assigned to t. In summary, the first stage finds the best matching of subject cells to the cells of each template based on minimizing Euclidean distances, and the second stage finds the best matching of subject cells to a *label* by, in effect maximizing the number of template cells that support the assignment. Because the bipartite matching minimizes a global cost and guarantees a one-to-one mapping, it is superior to a simple majority vote scheme. Note that the result does not depend on the processing order of the templates.

3. Imposing AP/LR/DV Constraints

The bipartite matching scheme in §2 does not consider the relative spatial relationship among vertices within V^S or within V^T. For example suppose a pair of cells (a,b) in the subject S should be mapped to a pair of cells (u,v) in the template T, with a to u, b to v, where it is always the case that u is anterior to v in *all* the templates. The unconstrained bipartite matching is free to match a to v and b to u and this is likely wrong. To solve this problem, we propose using invariant anterior-posterior (AP), left-right (LR), and dorsal-ventral (DV) relationships between cells to prune the possible match edges, i.e. the set of edges E between V^S and V^T in the bipartite graph model.

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

3.1 Deriving the Intrinsic AP/LR/DV Relationships Between Cells

The intrinsic AP/LR/DV relationships among cells are derived from the template images. Let us take the AP relationship as an example. We compute the $|L|\times|L|$ adjacency matrix $\mathbf{AP}_k$ for each template T_k, where $\mathbf{AP}_k(u,v) = 1$ if cell u is anterior to cell v, or either of u or v is in $L - V^{T_k}$, and 0 otherwise. Then the consensus AP adjacency matrix, denoted **AP**, can be obtained by applying the simple element-wise AND operation, $\wedge$, on the $\mathbf{AP}_k$, i.e., $\mathbf{AP} = \mathbf{AP}_1 \wedge \mathbf{AP}_2 \wedge \ldots \wedge \mathbf{AP}_K$. In this matrix, $\mathbf{AP}(u,v) = 1$ if and only if cell u is always anterior to cell v in all K templates, and 0 otherwise (we are assuming that every label is used in at least one template). In the same way, we also compute the LR/DV adjacency matrix **LR** and **DV** to describe the intrinsic LR and DV relationships among cells across different images.

3.2 Constructing AP/LR/DV Adjacency Matrices for a Subject Image

Given a matching M that maps cells in the subject S to cells in a template T_k, we may construct AP/LR/DV adjacency matrices for S, denoted **ap**, **lr**, and **dv,** as follows. If a pair of cells a and b in S are recognized as cells u and v in T_k, respectively, under the bipartite matching M, i.e., $a \rightarrow u \in M$ and $b \rightarrow v \in M$, and cell a is anterior to b in the subject image then we set $\mathbf{ap}(u,v) = 1$. We also set $\mathbf{ap}(u,v) = 1$ if u or v is in $L - V^{T_k}$. Otherwise $\mathbf{ap}(u,v) = 0$. Similarly, we compute the LR/DV adjacency matrices **lr** and **dv**. In brief, the spatial relationships of the subject are mapped to the template via the matching M.

3.3 Selecting Wrongly Recognized Cells and Pruning Impossible Edges of the Bipartite Graph

Given $a \rightarrow u \in M$ and $b \rightarrow v \in M$, if $\mathbf{ap}(u,v)=1$ and $\mathbf{ap}(v,u)=0$, but $\mathbf{AP}(u,v)=0$ and $\mathbf{AP}(v,u)=1$, then it is the case that cells a and b in the subject, where a is anterior to b, are labeled as cells u and v, with u always posterior to v in the templates they occur in. Thus at least one of the cells a and b in the subject is matched incorrectly. More generally, we may compute a contradiction matrix **C** using the 6 adjacency-matrices:

$$\mathbf{C} = \mathbf{C}_{\mathbf{ap}} \vee \mathbf{C}_{\mathbf{lr}} \vee \mathbf{C}_{\mathbf{dv}} \tag{4}$$

$$\mathbf{C}_{\mathbf{r}} = [(\mathbf{R}) \wedge (\neg \mathbf{R}^{\mathbf{T}}) \wedge (\neg \mathbf{r}) \wedge (\mathbf{r}^{\mathbf{T}})] \vee [(\neg \mathbf{R}) \wedge (\mathbf{R}^{\mathbf{T}}) \wedge (\mathbf{r}) \wedge (\neg \mathbf{r}^{\mathbf{T}})] \tag{5}$$

where $\vee, \wedge, \neg$ are the element-wise OR, AND, and NOT operations, respectively, and $^{\mathbf{T}}$ is matrix transposition. **R** represents adjacency matrices AP, LR, DV, and **r** represents adjacency matrices ap, lr, dv respectively. Moreover, when **r** equals say **ap** in Eq. (5) then **R** is **AP**. Observe that $\mathbf{C}(u,v) = 1$ if and only if one or more of the AP, LR, or DV relationships of cells a and b in the subject image are contradictory to those of cells u and v in the template. Thus at least one of a and b is wrongly recognized.

Based on the contradiction matrix **C**, we select, with high confidence, the cells in the subject that are wrongly labeled by M. We then cut the edges between these cells in the subject image and their mappings in the template and rerun the bipartite matching. To select the cells that are most likely to be wrongly recognized, we count, for each cell a in the subject S, the number of cells in S that have a contradictory AP/LR/DV relationships with cell a, i.e.,

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

$$conflict(a) = |\{b \mid \mathbf{C}(u,v) = 1, a \rightarrow u \in M, b \rightarrow v \in M\}| \quad (6)$$

We then take the most conflicted cell and remove the edge between it and its assigned vertex in M. We then compute using the Hungarian algorithm [15] a new M with respect to the reduced bipartite graph. This process is repeated until $\Sigma_a\, conflict(a)$ does not decrease for t_{max} sequential steps (t_{max}=3 for the results reported, but other values yielded similar results.). Once terminated, one takes as the answer the matching M that gives the minimum $\Sigma_a\, conflict(a)$.

We have thus far not identified M as M^* or one of the M_k. We actually find conflicts for each stage 1 matching M_k and should technically speak of $\boldsymbol{C}_k$. That is, we produce the best subject to template matching for each template using the matrices **AP**, **DV**, and **LR** that represent the invariant relationships over all the templates. Thereafter, we proceed with compute M^* in stage 2 as before. Algorithm 1 in Appendix shows the pseudo-code of the AP/LR/DV constrained two-stage bipartite matching.

4. Marker-based Recognition

The recognition approach above treats each cell equally and matches them all together at once. However, biologists usually use markers to aid cell identification. For instance, in the manual annotation of cells in *C. elegans*, our biologists first assigned names to the body wall muscle cells that were stained separately with GFP. With these marker cells labeled, the biologists then annotated both the ventral motor neurons and intestinal cells by examining their positions relative to the marker cells. After that the biologists used relative triangulation to annotate most other cells in trunk. Therefore, in addition to AP/LR/DV-constrained bipartite matching, in the following we present a hierarchical strategy similar to that of the biologists by first identifying marker cells and then using these marker cells to aid the identification of other cells.

4.1 Recognition of Muscle Cells

In the L1 larval stage of *C. elegans*, there are 81 body wall muscle cells and 1 depressor cell distributed along the entire worm body from the head to the tail. In our data (see §5 for details), most muscle cells, lit up by GFP in a separate frequency channel, are well separated from each other and thus are easier to segment and recognize, compared to other cells. We thus first use the AP/LR/DV-constrained bipartite matching to recognize just these 82 cells in the GFP channel. In this case, the adjacency matrices, **AP**, **LR**, **DV**, **ap**, **lr**, **dv** and an assignment are computed only for these 82 cells.

4.2 Identifications of Additional Markers

Once the identities of 81 body wall muscle cells and the 1 depressor cell have been determined, we use them as markers to identify cells that can be uniquely determined according to their relative positions with respect to these muscle cells. For this purpose, we again make use of adjacency matrices of template images and of the subject image. However, at this stage we only care about the relative relationship of cells to be recognized with respect to the markers. Thus we use sub-matrices of the six adjacency matrices. Using **AP** as an example, we extract 2 sub-matrices, denoted as

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

$\mathbf{AP}_{(pxq)}$ and $\mathbf{AP}_{(qxp)}$. The sub-matrix $\mathbf{AP}_{(pxq)}$ contains p rows and q columns. The p rows correspond to the p cells in the template to be matched by the cells in the subject image. The q columns correspond to the q marker cells (i.e., 82 in this example). Note that $p+q = N^T$ is the total number of cells annotated in the template images. The sub-matrix $\mathbf{AP}_{(qxp)}$ contains q rows and p columns, corresponding to q marker cells and p cells to be matched by cells in the subject image. The combination of $\mathbf{AP}_{(pxq)}$ and $\mathbf{AP}_{(qxp)}$ reflects the relative AP relationship between a cell and a marker. More specifically, if we denote $\mathbf{B_1}=\mathbf{AP}_{(pxq)} \wedge (\neg \mathbf{AP}_{(qxp)}^{\mathbf{T}})$, and $\mathbf{B_2}=(\neg \mathbf{AP}_{(pxq)}) \wedge (\mathbf{AP}_{(qxp)}^{\mathbf{T}})$, then cell u is anterior to marker v if $\mathbf{B_1}(u,v)=1$ and $\mathbf{B_2}(v,u)=0$, posterior to marker v if $\mathbf{B_1}(u,v)=0$ and $\mathbf{B_2}(v,u) = 1$, and can be either posterior or anterior to marker v if $\mathbf{B_1}(u,v)=0$ and $\mathbf{B_2}(v,u)=0$. Similarly, we compute $\mathbf{LR}_{(pxq)}$, $\mathbf{LR}_{(qxp)}$, $\mathbf{DV}_{(pxq)}$ and $\mathbf{DV}_{(qxp)}$.

We also extract the sub-matrices of **ap**, **lr**, and **dv**, denoted as $\mathbf{ap}_{(rxq)}$, $\mathbf{ap}_{(qxr)}$, $\mathbf{lr}_{(rxq)}$, $\mathbf{lr}_{(qxr)}$, $\mathbf{dv}_{(rxq)}$ and $\mathbf{dv}_{(qxr)}$. The r rows (in $\mathbf{ap}_{(rxq)}$, $\mathbf{lr}_{(rxq)}$, $\mathbf{dv}_{(rxq)}$) or r columns (in $\mathbf{ap}_{(qxr)}$, $\mathbf{lr}_{(qxr)}$, and $\mathbf{dv}_{(qxr)}$) correspond to the r cells in the subject image to be recognized (note that $r \geq p$). The q columns (in $\mathbf{ap}_{(rxq)}$, $\mathbf{lr}_{(rxq)}$, $\mathbf{dv}_{(rxq)}$) or q rows (in $\mathbf{ap}_{(qxr)}$, $\mathbf{lr}_{(qxr)}$, and $\mathbf{dv}_{(qxr)}$) correspond to the q cells in the subject image that have been recognized as markers (i.e, 82 muscle cells in this example). Note that $r+q = N^S$ is the total number of segmented cells in the subject image S. With these adjacency sub-matrices available, we further derive three matrices:

$$\begin{aligned}\mathbf{H}^{(\mathbf{r})}_{(r\times p)} &= [h^{(\mathbf{r})}]_{(r\times p)} = [(\mathbf{r}_{(r\times q)}) \wedge (\neg \mathbf{r}_{(q\times r)})^{\mathbf{T}}] \times [(\neg \mathbf{R}_{(p\times q)})^{\mathbf{T}} \wedge (\mathbf{R}_{(q\times p)})] \\ &+ [(\neg \mathbf{r}_{(r\times q)}) \wedge (\mathbf{r}_{(q\times r)})^{\mathbf{T}}] \times [(\mathbf{R}_{(p\times q)})^{\mathbf{T}} \wedge (\neg \mathbf{R}_{(q\times p)})]\end{aligned} \tag{7}$$

where $\times$ is matrix multiplication operation. **R** represents adjacency matrices **AP**, **LR**, **DV**, and **r** represents adjacency matrices **ap**, **lr**, **dv** respectively, similar to Eq. (5).

We then binarize $\mathbf{H}^{(\mathbf{r})}_{(r\times p)}$, obtaining $\mathbf{C}^{(\mathbf{r})}_{(r\times p)}$:

$$\mathbf{C}^{(\mathbf{r})}_{(r\times p)} = [c^{(\mathbf{r})}(a,u)]_{(r\times p)} = \begin{cases} 1 \text{ if } h^{(\mathbf{r})}(a,u) > 0 \\ 0 \text{ if } h^{(\mathbf{r})}(a,u) = 0 \end{cases} \tag{8}$$

If an element $c^{(\mathbf{ap})}(a,u)$ in $\mathbf{C}^{(\mathbf{ap})}_{(r\times p)}$ is 1, then the AP relationships between cell a and the marker cells in the subject image are different from those between cell u and the marker cells in the template. Thus cell a should not be recognized as cell u. The edge between a and u in the bipartite graph should be cut. On the contrary, if $c^{(\mathbf{ap})}(a,u)=0$, then the AP relationships between cell a and the marker cells in the subject image are consistent with those between cell u and the markers in the template. Thus cell a can be recognized as cell u. The edge between a and u in the bipartite graph should be kept. Similar explanation applies to $c^{(\mathbf{dv})}(a,u)$ and $c^{(\mathbf{lr})}(a,u)$.

Considering AP, LR, DV relationships all together, the contradictory matrix is computed as

$$\mathbf{C}_{(r\times p)} = \mathbf{C}^{(\mathbf{ap})}_{(r\times p)} \vee \mathbf{C}^{(\mathbf{lr})}_{(r\times p)} \vee \mathbf{C}^{(\mathbf{dv})}_{(r\times p)} \tag{9}$$

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

We then search for pair-wise cells (a,u) in matrix $\mathbf{C}_{(rxp)}$, such that $\mathbf{C}(a,u) = 0$, and $\forall x \neq u$, $\mathbf{C}(a,x)=1$, $\forall x \neq a$, $\mathbf{C}(x,u)=1$. This condition means cell a in the subject image can only be recognized as cell u in the template and at the same time cell u can only be assigned to cell a. In another word, cell a can be uniquely identified based on its relative position with respect to the markers. Cells thus identified are added to the set of markers. For those pair-wise cells (a,u) such that $\mathbf{C}(a,u) = 1$, we cut the edge between them in the bipartite graph by setting the distance between a and u to infinity.

After expanding the marker set, we repeat the above process until no new marker cell can be found. The remaining cells that cannot be uniquely determined according to their relative relationship with respect to markers are then recognized using AP/LR/DV constrained bipartite matching as described in §3. Algorithm 2 in Appendix shows the pseudocode of the marker-guided, AP/LR/DV constrained, two-stage bipartite matching.

5. Experiments

We applied our ARC method to the 3D images of newly hatched first larval stage hermaphrodites of *C. elegans* that were acquired using a Leica confocal microscope with 63x/1.4 oil lens. We used DAPI to stain the nuclei of all 558 cells, and GFP to stain the nuclei of the 81 body wall muscle cells and 1 depressor muscle cell (see Figure 1 for example data). As a worm body usually curves in 3D, we developed an automated approach to straighten a curved worm body into a canonical rod shape to facilitate later image comparison across different individuals [3] (see example in Figure 1 (b)). We then segmented cells in 3D using adaptive thresholding, the watershed algorithm, and a region grouping method (the details of the method [2] are beyond the scope of this paper, thus they are omitted). After that, we normalized each worm image, making the sizes and the orientations of different worms the same. This step maps the coordinates of the cells into a standard space defined by AP, LR, and DV axes. Finally, we annotated cells in a set of images with the aid of a 3D annotation tool called WANO developed by us (see Figure 1 (b) for schematic example of 3D annotated templates). Cells in the nerve ring of the head are small and tightly clustered and so very difficult to annotate solely based on our current images without developmental or cell-specific staining information. We annotated the subset of 351 cells out of the 558 cells, that exclude most neurons in the pharynx. These annotated cells include all the body wall muscle cells distributed along the entire worm body, 99 cells in the trunk where cell densities are relatively low, and 170 additional cells of different types in the head and tail. Thus our purpose was to match a subset of ~558 segmented regions in a subject image against the 351 annotated cells in templates.

One of the key ideas in this paper is to use the relative location relationships among cells to constrain the possible matching. We computed and analyzed the AP, DV, LR adjacency matrices from template images. Figure 2 illustrates the invariant AP relationship. For clarity of displaying, we only show the AP relationships of 181 cells by plotting the AP adjacency matrix as a graph after transitive reduction. It can be seen that many cells have strong AP relationships, apparently due to the stereotypy of *C. elegans* cells. These relationships, as well as the DV and LR relationships were used to constrain the possible mappings between cells in a subject image and the templates.

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

We used 15 image stacks and leave-one-out cross validation scheme to test our recognition method. In other words, we repeated the experiment 15 times. Each time we took one image as the subject image and the remaining 14 as the template images. Our purpose was to identify from all the segmented cells in each subject image the 351 cells that had been annotated in the templates. We compared our three approaches: two-stage bipartite matching (BM), AP/LR/DV constrained two-stage bipartite matching, and marker guided AP/LR/DV constrained two-stage bipartite matching.

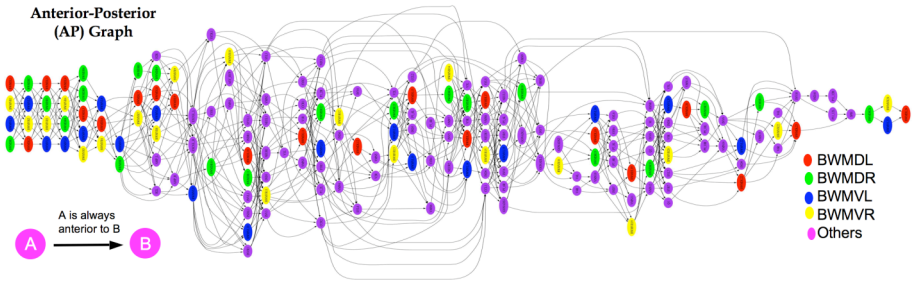

Fig. 2. Illustration of the invariant AP relationship of cells. For clarity of visualization, only the transitive reduction of the AP adjacency matrix is shown here for a set of 181 cells, including all 82 muscle cell markers and all cells in the trunk of L1-stage *C. elegans*. In this figure, left is anterior and right is posterior. An arrow always points from left to right (i.e. anterior to posterior).

The results in Table 1 show that using only the spatial coordinates of cells without considering the relative relationships between cells, the bipartite matching can only achieve an average of 73.79% accuracy in recognizing the 351 cells from the ~558 segmented regions (the second column in Table 1). When adding AP/LR/DV constraints to tailor edges in the bipartite graph, the accuracy improved ~5% (the third column in Table 1). With the combined use of marker-guides and AP/LR/DV-constrained bipartite matching, the average recognition accuracy improved significantly to 94.91% (the fourth column in Table 1). In this case, the average recognition rate of muscle cells (markers) is 99.81% (all 100% except that for stack S_{13}, which is 97.56%) (not shown in Table 1). Thus the average recognition rate of the remaining 269 cells is 93.42%. This indicates that the accuracy improvement is not merely due to the increased number of muscle cells that are correctly recognized in a separate channel but due to the marker guided scheme.

We also compared our method against other conventional approaches such as the K-Nearest-Neighbor (KNN) classifier and soft assignment approach [7]. The KNN approach finds for each cell in the subject image the K closest cells in the templates and then use majority vote to determine cell identities. The method did not yield a leave-one-out accuracy higher than 60%, much lower than our results showed in Table 1. The soft assignment method is computationally very expensive for big graphs. Thus we tested the recognition of 99 trunk cells. Despite the low number and low density of these cells which makes the task easier than our original matching problem, our results show that the average recognition rate using soft assignment is no higher than 68%.

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

We further analyzed for each cell to be recognized, how many times in the 15 images it is wrongly recognized. We then computed distribution of the cells and plotted the percentage of cells as a function of the number of images in which a cell was wrongly recognized. The result is shown in Figure 3. Among the 351 cells, 71% (the left most bar) of them are correctly recognized in all the 15 images, 90% (the sum of the three left most bars) are correctly recognized in 13 to 15 of the15 images. In the worst cases, there are two cells wrongly recognized in 7 images and another two cells wrongly recognized in 9 images (the two right-most bars). Those cells do not have a fixed local spatial relationship with respect to their neighboring cells and are in the head where cells are more densely clustered in the animal.

Table 1. Comparison of the recognition rates of the two-stage bipartite matching (BM), AP/LR/DV constrained two-stage BM, and marker-guided-AP/LR/DV-constrained-two-stage BM. The rates are produced by leave-one-out cross validation on 15 image stacks.

Image stack	Two-stage BM	AP/LR/DV constrained BM	Marker guided AP/LR/DV constrained BM
S_1	0.7114	0.7771	0.9486
S_2	0.7593	0.8166	0.9341
S_3	0.7607	0.8319	0.9829
S_4	0.7721	0.8205	0.9288
S_5	0.7892	0.8689	0.9259
S_6	0.5244	0.6074	0.9799
S_7	0.8054	0.8084	0.9581
S_8	0.7216	0.7994	0.9731
S_9	0.6161	0.6726	0.9821
S_{10}	0.9017	0.9153	0.9186
S_{11}	0.8328	0.8396	0.9147
S_{12}	0.6944	0.6458	0.9271
S_{13}	0.7550	0.8177	0.9459
S_{14}	0.7229	0.7971	0.9571
S_{15}	0.7009	0.8034	0.9601
mean	0.7379	0.7881	0.9491

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30-April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128-139, 2008.

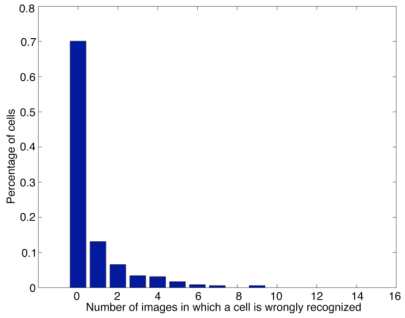

Fig. 3. The percentage of cells *P(k)* incorrectly recognized in *k* of the 15 images.

Overall, the experimental results show that our method can achieve high recognition accuracy despite the difficulty of the problem. To further improve the recognition accuracy, we will use additional cell information, such as size, shape, and gene expression levels. In fact, although our method currently only uses spatial coordinates and the relative spatial relationships between cells, our scheme is general enough to incorporate this additional cell information for further improvement.

Acknowledgement. The authors thank Andrew Fire for providing reagents and advice. Three-dimensional image stacks were generated in the Cell Sciences Imaging Facility of Stanford University. The authors acknowledge the financial support of the Larry L. Hillblom Foundation for XL. The work of XL and SK was funded by the Ellison Medical Foundation and the NIH.

Reference

1. Riddle, D., Blumenhal, T., Meyer, B., and Priess, J.R.: C. Elegans II, Cold Spring Harbor Laboratory Press, New York (1997).
2. Long, F., Peng, H., Liu, X., Kim, S., Myers, E.W.: A 3D digital cell atlas for the first larval stage of C. elegans hermaphrodite, HHMI JFRC Technical Report, 2007. Also Appear on 2007 Int. Conference of C. elegans, 2007.
3. Peng, H, Long, F, Liu, X., Kim, S., and Myers, E.: Straightening C.elegans Images. Bioinformatics (doi: 10.1093/bioinformatics/btm569 (in press) (2008).
4. Conte, D., Foggia, P., Sansone, C., and Vento, M., Thirty Years of Graph Matching in Pattern Recognition. International Journal of Pattern Recognition and Artificial Intelligence, 18 (3), 265-298 (2004).
5. Ullman, J.R.: An Algorithm for Subgraph Isomorphism. J. Assoc. Comput. Mach. 23, 31-42 (1976).
6. Llados, J., Marti, E., and Villanueva, J. J.: Symbol Recognition by Error-Tolerant Sub-Graph Matching Between Region Adjacency Graphs. IEEE Trans. Patt. Anal. Mach. Intell. 23, 1137-1143 (2001).
7. Gold, S., and Rangarajan, A.: A Graduated Assignment Algorithm for Graph Matching. IEEE Trans Patt. Anal. Mach. Intell., 18, 377-388 (1996).
8. Christmas, W. J., Kittler, J., and Petrou, M.: Structural Matching in Computer Vision Using Probabilistic Relaxation. IEEE Trans. Patt. Anal. Mach. Intell., 17(8), 749-764 (1995).

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30-April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128-139, 2008.

9. Umeyama, S.: An Eigendecomposition Approach to Weighted Graph Matching Problems. IEEE Trans. Patt. Anal. Mach. Intell., 10, 695-703 (1988).
10. Wilson, R. C., and Hancock, E.R.: Structural Matching by Discrete Relaxation. IEEE Trans. Patt. Anal. Mach. Intell., 19, 634-648 (1997).
11. Kotropoulos, C., Tefas, A., and Pitas, I.: Frontal Face Authentication Using Morphological Elastic Graph Matching. IEEE Trans. Imag. Process., 9, 555-560 (2000).
12. Chen, H. T., Lin, H., and Liu, T. L.: Multi-object Tracking Using Dynamical Graph Matching. Proc. 2001 IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition, 210-217 (2001).
13. Berretti, S., Del Bimbo, A., Vicario, E.: Efficient Matching and Indexing of Graph Models in Content-based Retrieval. IEEE Trans. Patt. Anal. Mach. Intell. 23, 1089-1105 (2001).
14. Ton J. and Jain, A. K.: Registering Landsat Images by Point Matching. IEEE Trans. Geoscience and Remote Sensing, 27(5), 642-651 (1989).
15. Munkres, J., Algorithms for the Assignment and Transportation Problems. Journal of the Society of Industrial and Applied Mathematics, 5(1), 32-38, (1957).

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30-April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128-139, 2008.

Appendix

Algorithm 1: Cell recognition using AP/LR/DV constrained two-stage bipartite matching
Input: A subject image S with N^S segmented cell regions and template images T_k, $k \in [1,K]$, with N^{T_k} annotated cells each, and a threshold t_{max}.
Output: Matching matrix M^* and cost value E^2

1. Compute the intrinsic adjacency matrices **AP**, **LR**, **DV**.
2. $\forall a, \forall u$, Set $N(a \Leftrightarrow u) = 0$
3. FOR EACH T_k
4. { Compute distance $D^1(a \rightarrow u)$ using Eq. (2)
5. Set $t = 0$, $minerr = \infty$
6. WHILE ($t < t_{max}$)
7. { Compute matchings M_k using the first stage bipartite matching
8. Compute adjacency matrices **ap**, **lr**, **dv** from the subject image S using M_k
9. Compute contradiction matrix **C** using Eqs. (4)~(5)
10. Select wrongly matched a and set $D^1(a \rightarrow u) = \infty$ for $a \rightarrow u \in M_k$
11. IF $\sum_a conflict(a) \le minerr$
12. $minerr = \sum_a conflict(a)$, $MB_k = M_k$
13. ELSE
14. $t = t+1$}
15. $M_k = MB_k$
16. $N(a \Leftrightarrow u) = N(a \Leftrightarrow u) + 1$, if $a \rightarrow u \in M_k$}
17. Compute $D^2(a \rightarrow u)$ using Eq. (3)
18. Compute the matching $\boldsymbol{M}^*$ and cost value E^2 using $D^2(a \rightarrow u)$ and the second stage bipartite matching

Algorithm 2: Cell recognition using marker guided, AP/LR/DV constrained, two-stage bipartite matching
Input: A subject image S with N^S segmented cell regions and K template images T_k, $k \in [1,K]$, each with N^{T_k} annotated cells, and a threshold t_{max}.
Output: Matching matrix M^* and cost value E^2

1. Compute adjacency matrices **AP**, **LR**, **DV** from template images T_k, $k \in [1,K]$.
2. Recognize muscle cells in the GFP channel by calling *Algorithm 1*.
3. Let U={all segmented regions in S}, U_m={recognized muscle cells in S}, V = {annotated cells in templates}, V_m = {annotated muscle cells in templates}
4. WHILE (new markers detected)
5. { $U = U \setminus U_m$, $r = |U|$, $V = V \setminus V_m$, $p = |V|$
6. Compute contradiction matrix $\mathbf{C}_{(r \times p)}$ using Eqs. (7)~(9)
7. Prune edges in the bipartite graph using $\mathbf{C}_{(r \times p)}$
8. Detect new markers, $U_m = U_m \cup$ {new markers in S}, $V_m = V_m \cup$ {new markers in T }
}
9. $\forall a \in U \setminus U_m$, $\forall u \in V \setminus V_m$, set $N(a \Leftrightarrow u) = 0$
10. FOREACH T_k
11. {Compute distance $D^1(a \rightarrow u)$ using Eq. (2)

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.

12. Set $t = 0$, $minerr = \infty$
13. WHILE ($t<t_{max}$)
14. { Compute matching matrix M_k using the first stage bipartite matching
15. Compute adjacency matrices **ap**, **lr**, **dv** from subject image S using M_k
16. Compute contradiction matrix **C** using Eqs. (4)~(5)
17. Select wrongly matched a's and set $D^1(a \rightarrow u) = \infty$ for $a \rightarrow u \in M_k$
18. IF $\sum_a conflict(a) \leq minerr$
19. $minerr = \sum_a conflict(a)$, $MB_k = M_k$
20. ELSE
21. $t = t+1$ }
22. $M_k = MB_k$
23. $N(a \Leftrightarrow u) = N(a \Leftrightarrow u) + 1$, if $a \rightarrow u \in M_k$}
24. Compute $D^2(a \rightarrow u)$ using Eq. (3)
25. Compute the matching $\boldsymbol{M}^*$ and cost value E^2 using $D^2(a \rightarrow u)$ and the second stage bipartite matching

Proc. of RECOMB 2008: 12th Annual Int. Conf. on Research in Computational Molecular Biology, March 30–April 2, 2008, Singapore; Lecture Notes in Computer Science: Research in Computational Molecular Biology, Springer Berlin / Heidelberg, pp. 128–139, 2008.