Eliminating Errors in Image Matching Caused by the Failure of Heuristics*

Gail Markovich and Michael Skolnick
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180

Abstract

An algorithm for determining correspondences between images has been developed and applied to the analysis of 2-D Electrophoretic gels[2, 6]. Various heuristics - both derived from the comparison problem and the specific application - are incorporated into the algorithm and result in a fairly high level of accuracy. Consistency checking mechanisms, based on exploring violations of a basic symmetry relation between correspondences, have been developed and incorporated into the algorithm to achieve higher accuracy[3]. The effect of these consistency checking mechanisms is to make inferences concerning the identification and cause of errors due to the (relatively rare) failures of heuristics used in earlier stages of the algorithm. Experiment analysis showing the benefit gained by this consistency methodology is presented.

Introduction

The automatic comparison of 2-D gels of proteins is being used in the detection of mutations[2, 6]. Mutations, because they are rare, add the requirement for high degrees of accuracy to the comparison problem, thus the need for a mechanism to detect and correct errors. Our algorithm addresses this problem by using a consistency checking mechanism which monitors a basic symmetry relation between matches. First, matches decisions are made based on a local measure of similarity. Then the relation between nearby matches are monitored with respect to a symmetry requirement. Procedures examine violations of this symmetry requirement which are then related back to the failure of the various heuristics, parameters and thresholds. We believe that this work provides a formulation of the image matching problem that is an alternative to the widely used relaxation procedure[1].

Comparison Algorithm

Two-dimensional Electrophoretic gels (2-D gels) are made by inducing a sample of a protein or DNA fragment of biological interest to migrate in two-dimensions on a polyacrylamide gel. A digitized image of a 2-D gel looks like a pattern of spots where each spot represents an object of biological interest and is characterized by a molecular weight and pH, see Figure 1. Comparison of 2-D gels involves detecting the correspondences between the spots from each gel and identifying spots with no correspondence.

Figure 1: An image of a 2-D DNA gel.

Before the 2-D gels images can be compared, both the locations and sizes of the spots on each gel image must be determined. Morphological algorithms have been used to perform this “low-level” processing of the gel images [5]. As in most low-level algorithms, a threshold operation is performed to distinguish the objects of interest from those objects judged to be insignificant. Associated with each possible spot location is a measure of the size of the spot. If the size measure is above some empirically chosen threshold then the spot is con-

*This research was supported in part by NIH grant NS21219 and USPHS grant S07RR07104. The 2-D gel application and data is part of a collaboration with Drs. Samir Hanash and James V. Neel of the University of Michigan.
Considered to be in a set of spots to be initially matched by the comparison algorithm. Thus, the input to the basic comparison algorithm is a set of points from each image divided into above and below threshold groups.

**Basic Algorithm**

The first step in comparing the sets of points in the above threshold group is to determine the neighbors of each point. The neighbors provide the edge relationship of the graph structure used in the matching algorithm. Several graph structures are possible. Figure 2a and 2b contain examples of neighborhoods from two gels being compared.

In the next step, the comparison algorithm determines a set of initial matches between the gel images, used on a comparison of the neighborhoods. A test on each pair of points, one from each image, is performed by asking a cross-correlational measure (match score). As illustrated in Figure 2c, potential matching spots (1001 and 1002 from Figures 2a and 2b) are centered at a common origin, and the degree of clustering in the relative positions of the neighbors of the match candidates determined by a cross-correlation measure. The relative locations of neighboring spots from the two gels are circled on Figure 2c when they “cluster” according to the cross-correlational measure. A positional noise function, applied to nearest neighbors from different gels, determines each cluster. The noise function determines the maximal allowable positional variation in each cluster as a function of the clusters distance from the origin. The number of clusters determines the candidates match score. A threshold on the match score determines the matches.

The control flow of the basic matching algorithm is used on a “committed” heuristic. Essentially this is a non-backtracking search strategy where the “clustered” pairs of a match become potential match hypotheses. The neighbors that cluster by a matched pair are considered match candidates, or “matches” (or quote matches) and are placed on a queue for terminating if they match by the correlation measure. The non-backtracking “committed” approach prohibits any point that has previously been queued from being placed on the queue again regardless of whether or not it has been matched previously or failed to match. An initial matching pair (seed match) must be found before proceeding. This is achieved by testing pairwise, spots from a small region on each gel, using the cross-correlational measure technique. The seed match chosen as the pair in the region with the highest measure, and is placed on the queue.

Generally, the set of “matches” in a matching pair, will not contain all the possible “matches”. Reasons for this are that 1. spots detected on one image may have a corresponding spot on the other image that was not detected due to the threshold used in processing the gel images, and 2. there may be genuine differences between the images, in which case there may be no point to participate in a match. For example, in Figure 2c, point 1041 is an isolated neighboring spot that formed no cluster. In order to determine whether such non-correspondences are real or not (reflect variations about a threshold or not) the image comparison algorithm searches the original image data for below-threshold spots that would reconcile the differences. If a below threshold spot is found, the newly found spot is added to the set of spots participating in the matching procedure. If no below-threshold spot is found, the comparison algorithm may create a “virtual” spot in its place (governed by a nearest neighbor heuristic in determining creation and placement). A virtual acts as a marker of differences between images. The newly added points can then participate in the matching process and can be considered as “match” hypotheses for further testing.

This basic matching algorithm results in a initial set of matches. Note the use of various heuristics (e.g., the graph structure, nearest neighbor bias in the clustering) thresholds (e.g., on the match score) and parameters (e.g., the noise function). As in all such heuristics, these tend to work as expected for the most part. These simple rules generate correspondences with an accuracy of about 80% for this application, at little computational cost. However, the use of these heuristics does introduce some uncertainty into the match decisions. As a result the set of matches produced is incomplete and has some errors particularly in locations of the gel that vary from the assumed heuristics and parameters. Therefore, a final consistency checking mechanism will be used to detect the relatively infrequent failures of the heuristics and parameters and invoke additional procedures to diagnose and correct these errors.

**Consistency**

The consistency checking mechanism involves monitoring a basic symmetry relation between matching pairs. Specifically, nearby pairs of matching spots are expected to contribute to each others correlational measure when the two pairs are consistent. For example, in Figure 2c the circled neighbors should become matches with the centered pair contributing to each of their correlational measures.

This pairwise symmetry is efficiently monitored between all pairs of matches using graph and matrix data structures as in Figures 2d and 2e. The nodes in this graph are labeled by pairs of spots (i,j) where node (i,j) is inserted into the graph if node i and node j from the original graphs were determined by the basic algorithm to match or “match” (cluster by a match). Note that if (i,j) is involved in both a match and “match” only one node is created, with the distinction between match and “match” embodied in the graph edges. Consistency is based on maintaining the following symmetric relationship between nodes in the “m”/“m” graph:

1. A pair is consistent with the current set of matches when all of the nodes adjacent to it have...
an edge coming back from it.

For example, in the graph of Figure 2d depicts the match of Figure 2c. Note that two of the pairs (1052,1051 and 1082,1071) which cluster in Figure 2c did not meet the symmetry requirement and are considered inconsistencies. This happens when for some reason to be determined latter by the algorithm - an independent cross-correlation of (1052,1051) and (1082,1071) did not occur or fell below the threshold on the match score.

Isolating and diagnosing inconsistencies involves a systematic exploration of violations in the symmetry of the edges of the graph. The method for identifying violations in symmetry begins by tabulating the counts of in and out edges for each node in the graph in to a matrix, as seen in Figure 2e. Each entry in the matrix contains two counts: the match score, which is the number of out edges, and the “match” score, which is the number of edges going into the node. The following two conditions characterize consistent symmetry between matches and “matches” in the matrix representation:

1. if the match score and “match” score of the i,jth entry of the matrix are equal in value (in degree = out degree for node (i,j)), and
2. if no other row or column coming out of i or j contains a non-zero “match” score (node labels are unique).

The submatrix in figure 2e corresponds to the portion of the graph in Figure 2d. Note the inconsistency in the graph appears in the matrix in rows 1001, 1031, 1071 and columns 1002, 1012, and 1082.

Violations of the symmetry relation are searched for in these data structures and the search gives rise to information on inconsistencies in the set of matches and provides alternative and possibly more consistent match pairs. That is, patterns of symmetry violations exist in the data structures that can be directly (and causally) related back to earlier failures of various heuristics. Identifying these patterns is the basis for additional computation that selectively focuses on the repair of the relatively small number of errors. For examples of these patterns and the procedures developed to diagnosing and correcting errors, see [3, 4].

Experiments and Results

See [2] and [6] for performance results with respect to the detection of mutations. Additional experiments were done in order to evaluate the consistency methodology. These tests demonstrate the ability of this methodology to identify errors generated by the use of the heuristics in the Basic Algorithm. In addition, results indicate that contextual information is incorporated into the matching process to disambiguate matches on regions of the gel containing significant image differences.

The experiments involve first examining the match scores between all pairs of possible matches on each gel and identifying possible matches by maximum match scores. This is an exhaustive procedure (computationally costly) to identify matches based only on a similarity measure of local regions (the match score). Next, these possible matches are compared with the matches resulting from the Basic Algorithm, revealing that the heuristic search of Basic Algorithm operates as well as the complete search 90% of the time. The 10% different from the Basic Algorithm are detected as errors by the consistency algorithm and are corrected.

Similarly, comparing again the matches determined by the match score only (which contain no consistency information), are compared with the matches after consistency checking procedures have been applied. On average, 80% of these matches are the same. Of the 20% that are different, 85% of the matches indicated by the consistency measure are the correct matches, compared to only 10% using just the match score. Upon closer examination, it is noted that the differences between the matches generated by these two procedures occur in regions of the gel pairs that have considerable differences. A formal analysis of this consistency methodology and a more direct comparison to other techniques used in image matching which enforce consistency, for example, relaxation, is underway.

References


