Comparison of DNA chip and Computer Vision Data

Technical Report RN/03/16

W. B. Langdon
Computer Science, University College, London
Gower Street, London, WC1E 6BT, UK

30 September 2003
Minor text revisions 29 Oct 2003

1 Introduction

There are some parallels to be drawn between the data used in computer vision applications and DNA or gene chip data increasingly used in Bioinformatics. So much so that, it has been suggested modern computer vision techniques (such as used in face recognition) might be good starting points for creating algorithms to process DNA chip data. In sections 2 and 3 we contrast the two types of data, while Sections 4 and 5 suggests ways DNA chip data might be rendered more easily interpreted.

2 Computer Vision Pixel data

The primary data used in computer vision and related applications are arrays of pixels (picture cells). While traditionally pixels are arranged in two dimensional (2D) arrays corresponding to a rectangular area of the computer monitor other arrangements are also used. For example, one dimension (1D) and three dimensional (3D) spaces (voxels). Time may be added as an additional dimension. For example, one may treat a video sequence as a 3D array of pixels. Higher dimensional arrays can also be used.

The data held in each array cell (i.e. describing each pixel) can also vary but common formats are binary (black or white), 8 bit (grey scale), 3×8 bit (3 colour), $n\times$float (n band satellite data).

Individual data values in pixels are strongly related. Pixel data values have strong neighbourhood relationships. (This is why it is common to store them in (potentially high dimension) arrays, so the array reflects the geometry of the data.) Pixel data is well understood and many array manipulations techniques have been developed which correspond to geometric transformations. Image processing often has access to strong domain knowledge, including reliable models of noise and estimates of typical noise level on individual data.

3 DNA Chip data

At present data is sparse. Since, in contrast to computer vision, few experiments have been performed and each experiment typically yields only ≈ 100 data frames.

While data is collected as a 2D image of light intensities, typically multiple pixels are pre-processed by the manufactures software to combine ≈ 16 spot intensities into a single pico-molar protein (gene product) concentration. These are presented as a list of floating point numbers.

3.1 DNA Chip noise

There are multiple sources of noise in DNA expression data. Starting with the most mechanical (and hence easiest to deal with) these are.

- Measurement of light intensity at each spot on the chip.

  – Well understood?

  – May be treated with conventional image processing techniques?

- Defects on the chip. These may be systematic defects in manufacture or scratches introduced by handling. Inhomogeneous application of the solution containing the mixture of DNA fragments to be measured.

  To a certain extent the DNA chips are designed specifically to cope with these errors. To reduce the effect of geometric inhomogeneity, spots of cDNA relating to the same gene can be widely scattered across the chip. Since each additional base in a cDNA strand requires at least four photolithography stages it is common to limit cDNA strands to 25 nucleotides. Since each protein will typically contain may more then 25/3 amino acids up to 16 different cDNA strands each matching a different part of the protein's DNA are used to uniquely identify the protein.

  Slightly different cDNA strands are bound to adjacent parts of the chip to indicate DNA strands which bind to the target spot but are not specific to it. I.e. they also bind to other cDNA (which, in theory, they should not). Such non-specific DNA is thought to give rise to "negative" expression values seen in some versions of pre-processing software[1] .

[1] Later versions of the software produce only positive values. But this appears to be just brushing the problem under the carpet.

> There may be a tendency to rely on the manufacturer's software or to discard troublesome data. However errors may be obvious by visual inspection of the raw image data.

The DNA chip manufacturer's software appears to be successful at giving a continuous indication of protein concentration with a linear relationship to the true concentration over a huge dynamic range (≈ 6 orders of magnitude). However there is always a threshold below which the biomolecules concentration cannot be reliably measured. Also above a certain concentration the chip saturates and no longer gives a linear response.

The most trouble some aspect of inferring protein concentration from gene chip measurements appears (to me) to be that the chips measure concentrations of DNA strands. It is necessary to use "reverse transcription" to convert proteins to DNA. This is a biological process. I expect it to be extremely noisy but worse to give rise to many systematic errors, since I expect the yield of a given "reverse transcription" solution to vary considerably for each protein in the solution.

4 Reformatting DNA Chip data

Note each DNA chip gives a continuous measurement per gene. The order of data in the list of measurement has no meaning. Each datum has meaning, while individual pixels do not.

B. F. Buxton argues that hard image processing problems can be solved by using a human to give the location of "landmarks" (such as centre pixel of the nose, each eye, the chin etc.) to an automatic system. Once this has been done for sufficient examples, automatic machine learning techniques can be use to build on the landmarks to give far more accurate image processing systems.

What are the landmarks in DNA chip data?

Much DNA chip data refers to genes with well known characteristics. (This is after all, why they were selected by the gene chip manufactures to have their expression (protein) concentrations measured.) In many cases (at least some of) the interactions or relationships between genes is known. Similarly, in many cases, the evolutionary history of genes is known. In some cases, we have estimates for when Nature invented particular genes. I.e. how old the genes are.

Biology is highly conservative. Many metabolic processes in a tissue (and hence proteins and hence genes) are extremely similar in many others. So much so that Biologists are often more interested in the differences than the similarities.

Can we start our search for "landmarks" by establishing a geometry to describe genes?

How far would a geometry based on placing genes next to each other based on their age get us?

If this does not capture the relationships sufficiently well, can we exploit the family tree of genes? I.e. use a geometry based on phylograms.

4.1 Automatic Reformatting

Due to costs etc. it is common for gene chip experiments to yield only a few hundred DNA chip measurements, leading to the familiar problem of having measurements of tens of thousands of variables (the genes) but very few samples for each. But, as mention above, I expect the vast majority of this data to be highly similar across experiments[2]. By combining (already published) data from multiple experiments one should be able to assemble many thousands of measurements for a large number of genes. Since Biology is so conservative, gene data could be taken from different species. However, perhaps, we should limit ourselves to mammalian samples.

From this large sample, we should be able to give a reliable picture of the typical gene landscape. I.e. what is so "obvious" that it is not being studied! Can this be used to give a "geometry" to gene chip data.

4.2 Knowledge Based Reformatting

Another idea is to use existing sources of knowledge or data to build a similar geometry as in the previous section. Of course another option is to combine both.

As mentioned above a lot is known about the genes. Much can be inferred from the now known DNA sequence of each gene.

5 First Steps

It is unclear how far this approach will take us, but a first step would be using various data visualisation techniques to radically change the way DNA chip data is presented to the Biologists. Age ordering and family trees are suggested as things to try first. However one will also have to persuade the users to concentrate upon the similarities in the data rather than the differences before one can hope they will be able to identify "landmarks".

[2] It is common practise to discard measurements of gene expression where the values do not change!