ProteomeGRID: towards a high-throughput Proteomics pipeline

Andrew W. Dowsey¹, Michael J. Dunn² and Guang-Zhong Yang¹

¹Royal Society / Wolfson Mic Lab, Imperial College London. ²Conway Institute of Biomolecular and Biomedical Research, University College Dublin.

INTRODUCTION

The Proteome is the genome blueprint of an organism. However, sequencing these genes does not provide enough data to develop new therapies. In fact, genes code the construction of proteins, the 'chemical building blocks' that give structure and function to living things. Cells have the same genome but differ by which genes are active and the corresponding proteins that are made. The proteome is complex – the human proteome contains approximately 5 million different proteins.

A more complete understanding of disease may be gained by looking at the proteins present within a diseased cell or tissue. Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) is the only method available capable of simultaneously separating thousands of proteins (figure 1). One of the key objectives of the biochemist is to identify the differential expression between control and experimental sample gels. That is, the protein spots that have been inhibited (dissipated), induced (appeared), or have changed abundance. Such information has enormous morphological and clinical potential e.g. disease identification, drug synthesis, proteomics.

Unfortunately 2-D PAGE exhibits much experimental variation. Geometric deformations in the protein pattern, due to gel inhomogeneities and current leakage, hampers the correct matching of corresponding spots, whilst non-uniform stain and overlaying spots contribute to uncertainties in expression quantification.

Until recently the bioinformatics techniques have been user-assisted spot detection and point pattern matching algorithms, thus requiring many hours of the biochemist's time per gel pair. The complexity of new automated algorithms [1] requires Grid infrastructure for the processing, archival, standardisation and matching algorithms, thus requiring many hours of the biochemist's time per gel pair. The complexity of new automated algorithms [1] requires Grid infrastructure for the processing, archival, standardisation and matching algorithms, thus requiring many hours of the biochemist's time per gel pair. The complexity of new automated algorithms [1] requires Grid infrastructure for the processing, archival, standardisation and matching algorithms, thus requiring many hours of the biochemist's time per gel pair. The complexity of new automated algorithms [1] requires Grid infrastructure for the processing, archival, standardisation and matching algorithms, thus requiring many hours of the biochemist's time per gel pair.

ProteomeGRID

The main objective of ProteomeGRID is to provide a Grid enabled high-throughput proteomics pipeline, encompassing 2-DE acquisition and differential protein expression analysis, through to robotic excision of interesting protein spots. The system provides the upstream automation for current efforts in large-scale mass spectrometric protein identification. To realise this goal, it is vital to provide a generalised framework that can benefit the entire proteomics community as a whole.

ProteomeGRID builds on the proTurbo cluster image computing engine, with Grid-enabled versions of novel automatic 2-DE image analysis algorithms, and server middleware based upon the Open Services Grid Architecture (OSGA - http://www.globus.org/). ProteomeGRID is detailed in figure 3.

HIGH-THROUGHPUT WITH proTurbo

The first step towards this goal is to overcome the computational and communications burden entailed by the image analysis of 2-DE gels, with OSGA enabled cluster computing. We have developed the high-throughput proTurbo framework, which utilises Condor (http://www.cs.wisc.edu/condor/) cluster management, with JESP/LSL imageless image compression, for task farming massive batches of 2-DE gels. Spawning has been tuned to office machines when their users are away, and a novel probabilistic scheduler has been developed to maintain high throughput in response to the likelihood of the owner returning.

Our results [3] show a 4.1 fold and 9.1 fold compression ratio, and no network overhead did not affect other users. With 40 workers a 32-speedup was seen, resulting in 85% resource efficiency (Figure 2a), and the eager scheduler reduced the impact of eVoids by 56% (Figure 2b).

Large-scale Statistical Expression Analysis (SEA)

In traditional techniques, symbolic representation of spots is determined at the very early stages of the processing pipeline. Once such a description is reached intrinsic errors, dependant on the spot modelling and matching algorithms, and the errorpropagation throughout the subsequent processing steps. In practice, if the analyses were modelled statistically and performed simultaneously between sets of technical replicates, then small significant expression changes over one pair could become significant when reinforced by the same consistent changes in the others. This permits the creation of a database of probabilistic baselines, or 'norms', which characterises confidence levels for protein expression under different experimental settings. Such evidence will build up a statistical formation model of 2-DE, which would then be mined to discover intrinsic trends, increase the acuity of new results and provide valuable feedback to 2-DE scientists on the sensitivity of their experiments. We call this Statistical Expression Analysis (SEA).

ADVANCED IMAGE NORMALISATION

Raw gels submitted into ProteomeGRID (Figure 4a) must first be calibrated, so that they can be analysed directly. To maintain high-throughput this stage must be fully automatic, and so we based our approach on our MR algorithm [1], which is fully image based – it requires no recourse to detecting spots (Figure 3b-e).

This has been enhanced with the following:

• 2-DE exhibits physical deformation - prevalent warping algorithms are designed to correct or induce optical distortion, and so cause protein over-expression under dilution, and under-expression during contraction (Figure 4a-c). To preserve volume-invariance, the intensity of each transformed pixel must be normalised by the Jacobian of the mapping at that point.

• Excluding artefacts, the volume and distribution of protein in the sample and references gets should be the same. Each parameter in the warping should affect the same volume of protein. Currently we subdivide the gels recursively (Figure 3d) by their centre of mass (Figure 4b), as a fast approximation.

• The major remaining systematic artefact is the inaccuracy in quantification due to staining, loading and focusing errors [4]. We compensate for these anomalies by fitting a piecewise B-spline basis field that maximises the similarity between the gels, every time the registration picks a new warping (Figure 4f).

REFERENCES