Reprinted from Trends in Analytical Chemistry – *February, 1982* Vol. 1, No. 6, pp. 131 – 135

Towards a complete catalog of human proteins

A comprehensive protein database is essential to our understanding of cell function. Two dimensional electrophoresis is the only method at present capable of providing the information necessary for the formation of such a database.

N. Leigh Anderson
Argonne, IL, U.S.A.

One of the major goals of modern biology is to understand the way in which a 'higher' cell (e.g. a human cell) works. This sort of understanding would involve compiling a comprehensive list of the functional parts of the cell (primarily proteins) and their corresponding genes and a basic knowledge of how these parts interact to produce the biochemical behavior characteristic of living cells. With respect to the extent and variety of current biological research, it is interesting to see how close we are to achieving this goal and to find out which technical developments are being used to do so.

Current shortcomings

A simple comparison of the number of human proteins that have been characterized in some detail (several hundred) with the number thought to be expressed in a single cell type (5–10,000) or in all the cell types throughout human development (a total of perhaps 30–50,000) indicates that the cataloguing of proteins has barely begun[1]. If one makes the plausible assumption that the unknown human proteins are as important as those currently known, it would appear that 90–99% of the functional content of human cells remains a mystery. Therefore, by any standards our information is woefully incomplete; medieval alchemists correctly identified a larger fraction of the elements than we have proteins and yet still failed to unravel the principles of chemistry. The number of proteins involved, while much larger than those of the elements, is not in itself staggering – in fact, there are more different parts in a Boeing 747 than proteins in man. However, the problem with such complexity in a biological context is that it is difficult to find the parts in the first place.

If one makes the plausible assumption that the unknown human proteins are as important as those currently known, then it would appear that 90–99% of the functional content of human cells remains a mystery.

Perhaps the major obstacle in the way of a more complete understanding of the cell is the lack of an adequate framework within which to carry out a systematic exploration of the proteins. Such a framework would need to have: (a) a means of uniquely identifying each entity; and (b) a means of ensuring (within limits) that all such entities can be systematically found (i.e. a guarantee that the information is complete). At present, individual proteins can be isolated and named, and their molecular weights, chemical properties, and subcellular locations determined. However, this information is generally not enough to ensure that a single protein isolated in two different laboratories will be recognized as being the same. By sequencing amino acids one can uniquely identify proteins (since by sequencing one is virtually carrying out complete chemical specification), yet this type of data is rarely available because of the large effort involved in its determination. Progress towards the complete enumeration of the proteins is likewise difficult through the accepted biochemical approach, since in biochemistry one generally discovers proteins that are related to or interact with known proteins or metabolites, or follows them in their reactions within a chemical pathway. Thus, discovery tends to follow a logical or semilogical path from one protein to another and considerable effort is devoted to the characterization of each molecule as it is isolated. Naturally, there is no guarantee that the most interesting or important proteins will be discovered first, nor is there any means of achieving completion save by exhaustion. One can, by extrapolating from the rate of protein discovery over the last decades, arrive at quite discouraging estimates of the time required to achieve the comprehensive understanding we desire. These are the difficulties associated with an attack upon a complex field without the benefit of an adequate organizing framework.

An organizing framework

In contrast, taxonomy, the classification of species of living things, provides an historical example of the successful application of an organizing principle in biology. Beginning with the concept that all species can be placed on a giant family tree based on their degree of similarity or (as later interpreted) genetic relatedness, it was relatively easy to arrive at the conclusion that good rules of classification could lead to a much better understanding of the variety of life (which had until then seemed chaotic). This led quite naturally to the desire to find and classify *all* living things. Thus, a system developed in which discovery and debate could have free reign. Now, although numerous species remain to be found, the major outlines of biological evolution on earth have been laid out with considerable confidence. This groundwork allows contemporary biologists to address problems of

real substance, such as the mechanisms of evolution and the relationship of individual genes to the evolutionary fate of a species or individual. At present, there appears to be no dissension from the view that a complete exploration of all the species on earth was necessary and that it was highly productive.

. . . the organizing framework needed for protein cataloguing must be based on some technique capable of resolving and detecting thousands of proteins.

As noted above, the organizing framework needed for protein cataloguing must be based on some technique capable of resolving and detecting thousands of proteins. The classical methods of protein fractionation, such as precipitation, column chromatography, or starch electrophoresis, can generally resolve up to ten components. Using a variety of such methods applied in series, it is possible to isolate particular proteins from complex mixtures; however, the resolution of any one step is far too low for cataloguing. Fortunately, several acrylamide gel electrophoretic procedures can be made to resolve approximately 100 components, which, while still insufficient, offers the possibility of achieving the required resolution in a two-step system if the steps are sufficiently dissimilar. Beginning in 1975, O'Farrell and others[2–5] developed systems exploiting this possibility. Each used isoelectric focusing (IEF) in small diameter gel rods, followed by a perpendicular electrophoretic step carried out in slab gels in the presence of the detergent sodium dodecyl sulfate (SDS). Since IEF separates proteins according to their isoelectric point (a function of chemical composition) while SDS electrophoresis separates by polypeptide chain length (approximately

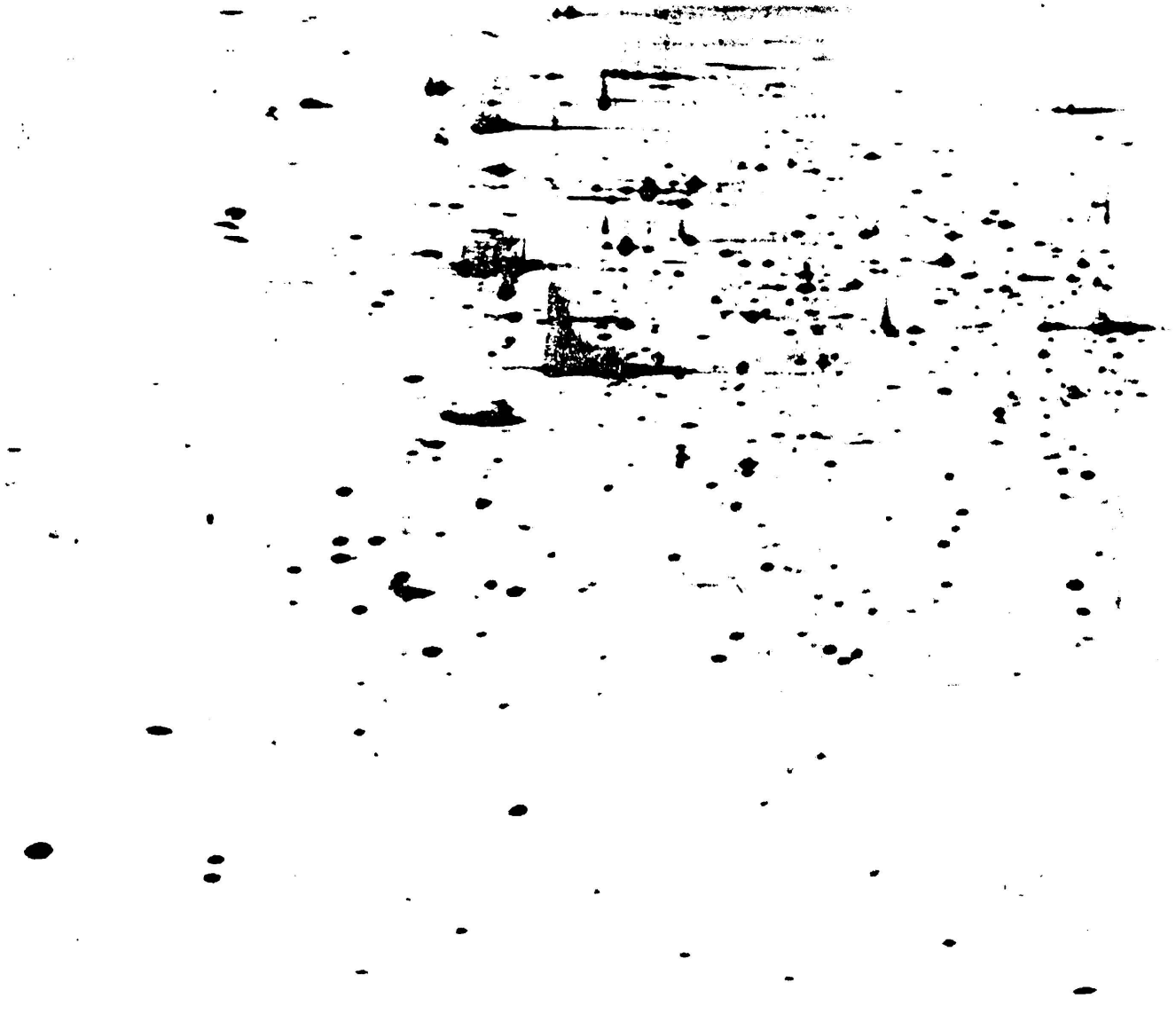

Fig. 1. Two-dimensional electrophoresis pattern of the proteins of a human lymphoblastoid cell line (GM607). The cells were grown in the presence of a radioactive amino acid ([^{35}S]methionine), then solubilized and subjected to two-dimensional electrophoresis. The 2-D gel was then dried and exposed to X-ray film to reveal the radiolabeled proteins. The isoelectric focusing dimension is horizontal (acid end to the left) and the SDS-electrophoresis dimension is vertical (with high molecular weight molecules at the top). Approximately 1000 distinct protein spots are detectable on the original autoradiogram.

"Medieval alchemists correctly identified a larger fraction of the elements and still failed to unravel the principles of chemistry".

equal to molecular weight), the two separations are effectively uncorrelated and the resulting resolution is about 100 × 100 = 10,000 proteins. O'Farrell's paper[2] demonstrated that it was possible in practice to resolve at least 1200 proteins from the bacterium *Escherichia coli*, a number close to the total number of proteins expected genetically. Samples of human cells can likewise be analysed to show 1000–2000 protein spots (Fig. 1). These results illustrate the potential usefulness of the technique in protein cataloguing and the feasibility of its use as a framework. Since no alternative technology has appeared to challenge it, extensive development has been undertaken to produce a standardizable, convenient separation system based on the O'Farrell technique.

. . . IEF separates proteins according to their isoelectric point, while SDS electrophoresis separates by polypeptide chain length, the two separations are effectively uncorrelated and the resulting resolution is about 100 × 100 = 10,000 proteins.

Maximum reproducibility is obtained within a batch of gels run simultaneously and, therefore, apparatus has been designed to run 10 or 20 samples together through both dimensions[6]. Comparability among different laboratories remains a problem because of the variation in chemicals from different suppliers, lot-to-lot variability of certain reagents such as the IEF ampholytes, and the use of various types of equipment having slightly different geometries. Nevertheless, the use of a finely calibrated protein marker series, in each dimension, allows a protein's position to be specified to within ~0.02 pH unit (IEF) and ~1,000 SDS-molecular weight units. This accuracy is generally sufficient to uniquely identify a protein in the two-dimensional patterns produced by different laboratories. The labor involved in performing two-dimensional analyses has been decreased to a point at which a modest gel laboratory can analyse 10,000 samples per year. This makes it possible to use the technique in large-scale genetic and toxicological screening experiments, as well as in basic biological research.

These developments make it reasonable to contemplate the use of two-dimensional electrophoretic analysis to systematically catalog human proteins and establish a database of information concerning them. Since most of the detectable proteins have unknown functions, the position of each protein spot in the two-dimensional map must identify it and spot serial numbers (attached to these positions) must serve as protein names. A wide variety of selective labeling, cell fractionation, and physicochemical techniques may then be applied to yield information about the proteins *en masse*. Labeling cells with $^{32}PO_4$, for instance, causes the phosphoproteins to become radioactive and permits their detection by autoradiography. Isolation of cell nuclei effectively enriches those proteins specific to the nucleus. By using two-dimensional analysis as the final step in protein detection (i.e. following a variety of other experimental techniques) all the information relating to a specific protein can be correlated. Thus, it becomes possible to know that protein 154 is a cytoskeletal phosphoprotein whose properties change following cell transformation, without ever having isolated it and without ever knowing its function. Of course, as information accumulates, the peculiar features of many proteins begin to suggest their possible functions and to hint at possible interrelationships.

Towards a protein database

The size of such a database (termed, in this case, the Human Protein Index) makes it unwieldy if committed to paper. In addition, it is essential to be able to manipulate the data easily. The solution to this problem is to computerize the database, provided that the two-dimensional patterns themselves can be computer analysed to yield an accurate list of spot positions and abundance. Several groups have recently constructed computer systems capable of achieving the latter goal[7–10], and it is now possible to quantitate 1000 proteins in a matter of minutes from the digitized image of a gel. By accurately matching the spots in each new analysis with a numbered reference pattern,

quantitative data can be integrated with the database and compared with the results of other analyses. The database must also include much more than quantitative information from gels; it must include, in an intelligible form, the conclusions to be drawn about the properties and possible functions of each protein. Because of the capabilities inherent in a computer-based, cross-referenced database, it then becomes possible to ask questions having to do with sets or types of proteins and how these change with respect to one another in complex situations.

In a larger sphere, a protein database – a Human Protein Index – may represent the only means available for assembling enough information on proteins to begin piecing together a comprehensive picture of normal cell function and its disruption in disease.

Clearly the ultimate usefulness of such a database depends on its widespread use and a general acceptance of the separation technology to which it is keyed. The only technology providing the required resolution today is two-dimensional electrophoresis, though in the future other high-resolution techniques might well take over. The increasing level of commercial interest in two-dimensional electrophoresis, in terms of reagents, gel apparatus, and computer processing, indicate that the technique will become progressively better, more convenient, and consequently more widely used. If clinical applications prove as promising as expected, a complete clinical system (including premade gels) may be produced, providing the research community with a level of reproducibility and interlaboratory comparability that is presently unobtainable. The effect of these developments on the course of biological research is difficult to predict in detail, but it is to be hoped that the emergence of a readily accessible protein data base will lead to the rapid exploration of vast areas of cellular function about which we are almost totally ignorant. At the very least, it will provide a framework within which disparate types of information obtained by a large number of investigators can be integrated and accessed. In a larger sphere, a protein database – a Human Protein Index – may represent the only means available for assembling enough information on proteins to begin piecing together a comprehensive picture of normal cell function and its disruption in disease.

Acknowledgments

I wish to thank my colleagues in the Molecular Anatomy Program, in whose company these ideas have evolved, and the U.S. Department of Energy, which supported this work under contract No. W-31-109-ENG-38.

References

1 Anderson, N. G. and Anderson, N. L. (1979) *Behring Inst. Mitt.* 63, 169–210
2 O'Farrell, P. H. (1975) *J. Biol. Chem.* 250, 4007–4021
3 Klose, J. (1975) *Humangenetik* 26, 231–243
4 Scheele, G. A. (1975) *J. Biol. Chem.* 250, 5375–5385
5 Iborra, G. and Behler, J.-M. (1976) *Anal. Biochem.* 74, 503–511
6 Anderson, N. G. and Anderson, N. L. (1978) *Anal. Biochem.* 85, 331–340
7 Garrels, J. I. (1979) *J. Biol. Chem.* 254, 7961–7977
8 Bossinger, J., Miller, M. J., Vo, K.-P., Geiduschek, E. P. and Xuong, N.-H. (1979) *J. Biol. Chem.* 254, 7986–7998
9 Lester, E. P., Lenkin, P., Lipkin, L. and Cooper, H. L. (1980) *Clin. Chem.* 26, 1392–1402
10 Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. and Anderson, N. G. (1981) *Clin. Chem.* (in press)

N. Leigh Anderson received his B.A. in physics from Yale University (1971), and his Ph.D. in molecular biology from Cambridge University (1974) where he worked with M. F. Perutz on the structure and mechanism of human hemoglobin. Since 1975, Dr Anderson has worked with his father (Dr N. G. Anderson) on the development of high-resolution protein separation systems and their use in establishing a Human Protein Index, in the Division of Biological and Medical Research, Argonne National Laboratory, Argonne, IL 60439, U.S.A.