Ligand assays: from electrophoresis to miniaturized microarrays

ROGER P. EKINS

The main developments in the “ligand assay” field in which I have been involved are traced. These include the original development of “first generation” competitive assays relying on radiolabeled analyte markers; the development of the first “second generation”, noncompetitive (ultrasensitive) methods, which rely on the use of labeled (monoclonal) antibodies and high specific activity nonisotopic labels (leading to the transformation of the immunodiagnostic field in the 1980s); and the development of the first “third generation” miniaturized, chip-based, microarray methods, which permit the simultaneous ultrasensitive measurement of many analytes in the same small sample. The latter—applicable both to immunoassay and to DNA/RNA analysis—are likely to revolutionize the diagnostic and pharmaceutical fields in the next decade.

Ligand Assays: General Concepts

It seemed to me to be appropriate, in the first of the Edwin F. Ullman Award presentations, to overview some of the key developments in the “ligand assay” field since its origins in the mid-1950s, particularly those in which I myself have been considerably involved. However, in a historical account of restricted length, it is clearly impossible to do justice to the work of the many scientists who have contributed to the field; this presentation’s primary intention is therefore to identify and clarify only the main concepts that have driven methodological advances in this area since its origins and that underlie those occurring at the present time.

The term ligand assay is generally used to describe the class of assays that rely on observation of the product(s) of the binding reaction between the analyte and a specific binding reagent, the latter comprising—in the case of immunoassays—an antibody directed against antigenic sites on the analyte molecule. Early immunoassays depended on direct observation of the binding reaction product, this typically taking the form of a precipitate of the antibody-antigen complex. Likewise, many so-called “bioassays” have relied on the binding reaction between the analyte and specific receptors present in cells and tissues. However, the term ligand assay as generally understood, and as used in this article, is restricted to assays relying on high specific activity labels (e.g., radioisotopic labels) to directly reveal the products of the binding reaction, thereby (amongst other benefits) increasing assay sensitivity.

Antibodies are readily prepared against many substances of biological interest, and label-based immunoassays (such as radioimmunoassays) have achieved widespread popularity in the past 30 years. Nevertheless, they conform to analytical principles that govern a number of analogous ligand assays that differ only in the nature of the binding reagent on which they rely. These methods [which exploit the specific binding characteristics of serum and receptor binding proteins, enzymes, single-chain DNA fragments (i.e., oligonucleotide probes), and others] have not hitherto achieved as great prominence or industrial importance as immunoassays; however, burgeoning interest in DNA analysis suggests that oligonucleotide-based ligand assays will prove at least equally important in the future.

The three principal phases in the development of this field, and their origins, are identified and summarized in Fig. 1:

Phase 1, the independent development in the 1950s (by Yalow and Berson in New York and by me in London) of the first generation “competitive” radiolabeled analyte methods culminating, at the end of the decade, in the publication of the “radioimmunoassay” (RIA) for insulin (1) and of the “saturation assay” for thyroxine (2);

Phase 2a, the development [by Wide et al. (3) in Uppsala and Miles and Hales (4) in Cambridge] of radiolabeled antibody (“immunoradiometric”) methods in the late 1960s;

Phase 2b, the collaborative development (by myself and Wallac Oy, commencing in 1976) of the first of a
second generation of “ultrasensitive”, nonisotopic, non-competitive labeled monoclonal antibody methods, this being ultimately marketed as “DELFIA” (5); and

Phase 3, the development (by myself and my co-workers, commencing in 1986) of miniaturized, multianalyte “microspot” assays (6, 7), leading, in 1991–1992, to a collaboration with Boehringer Mannheim GmbH on the development of a third generation of ultrasensitive antibody and oligonucleotide microarray technologies. (Note: a similar long-term project—the Genosensor Project—was subsequently launched in the United States in 1992, albeit this is directed solely to DNA/RNA analysis.)

Since its origins, a principal factor driving methodological developments in this area has been the need for greater assay sensitivity and, concomitantly, a reduction in incubation and processing times (implying higher sample throughputs). In these respects, the earliest competitive radioligand assays clearly represented a major advance on methods for the assay of substances of biological importance available in the 1950s, permitting, for example, the determination of many hormones whose concentrations in body fluids ($\sim 10^{10}$–$10^{17}$ molecules/L, or $10^7$–$10^{14}$ molecules/mL) lay below the sensitivity limits of contemporary analytical techniques. Because of their specificity, sensitivity, and ease of use, these methods rapidly penetrated into many other areas, e.g., hematology and pharmacology, emerging as the dominant microanalytical methodology used in medicine from the 1960s to the mid-1980s.

Although (radio)labeled antibody (immunoradiometric) assays were first described in the late 1960s, doubts regarding their sensitivities relative to those of “radiolabeled analyte” methods, coupled with the difficulties in, and cost of, producing relatively pure labeled antibodies, led to the virtually universal neglect of this approach by researchers and manufacturers until the mid-1980s. Thus, even the description by Köhler and Milstein (8) in 1975 of the in vitro synthesis of monoclonal antibodies—a technique that removed many of the logistic problems associated with the large-scale production of pure, labeled antibodies—generated little immediate interest among researchers and kit manufacturers. Indeed, as late as 1982 and even later, a prominent scientist in the field publicly asserted that monoclonal antibodies offered no substantial benefits, implicitly failing to recognize the potential advantages (in terms of improved sensitivity and reduced incubation times) of the labeled antibody approach. These failures permitted, inter alia, the use of monoclonal anti-
bodies for immunoassay purposes to be successfully patented (in the United States) as late as 1980 [see (9)].

However, after the demonstration (theoretical and experimental) of the improved performance achievable using noncompetitive, labeled (monoclonal) antibody methods (particularly when accompanied by the use of high specific activity nonspecific labels), the immunodiagnostic field was transformed in the mid-1980s, and virtually all kit manufacturers subsequently adopted this approach. The resulting improvement in assay performance provided the basis for the explosive development of the fast, high-throughput immunoanalyzers that now dominate the field.

The third phase identified in Fig. 1 promises as great an ultimate transformation of the ligand assay field as followed the emergence of the ultrasensitive methods in the early 1980s. Although the earliest studies on the feasibility of an ultrasensitive microarray technology commenced in my laboratory as long ago as 1986, it represents a far greater challenge to manufacturers than posed by the earlier methods, and the development time and costs involved are inevitably greater. Miniaturized microarray technologies permit, in principle, the simultaneous, ultra-sensitive measurement of tens, hundreds, or even thousands of substances in a small sample (e.g., a drop or two of blood). Arrays of high capacity are clearly likely to prove especially valuable in the context of DNA analysis; however, very large analyte “menus” are unlikely to be required in the immunodiagnostic field in the immediate future (although if methods are developed that permit the isoforms of analytes of heterogeneous molecular composition to be individually distinguished, a comparable need in this area may ultimately arise). Nevertheless, even at the present time, miniaturized microarray technology offers substantial advantages in certain obvious immunodiagnostic applications (such as in transfusion blood screening and allergy testing); moreover, the technology, once fully developed, is likely to offer major cost advantages when the determination of more than one analyte is required and could, therefore, replace most current methods even in circumstances where analyte panels of relatively small size are of diagnostic value.

**On the Meaning of Assay “Sensitivity” and “Precision”**

As indicated above, the need for high sensitivity underlay the original inception of ligand assay methods and has constituted a major factor driving their ongoing development. However, notwithstanding this word’s long-established meaning as indicative of the ability of an assay system to determine (or “sense”) low analyte concentrations, differences regarding the concept it represents, and the way in which the sensitivity of an assay system should therefore be correctly assessed, led to early and prolonged dispute regarding RIA design. This topic has been discussed at length in recent publications (10–13) and hence need not be dwelt on in detail here. Nevertheless, it is appropriate to summarize the issues involved because misunderstandings regarding the assessment of assay performance have played a major part in the history of the ligand assay field.

Certain (although not all) relevant national and international organizations—including, for example, IUPAC—have, in recent years, formally defined the sensitivity of a measuring system in terms of the response/stimulus ratio or (equivalently) the slope of the dose–response curve (14). Of greater importance in the present context, Berson and Yalow [in their many publications relating to immunoassay theory, e.g., (15, 16)] likewise defined the sensitivity of an assay in terms of the slope of the curve relating the response variable [the bound-to-free ratio (B/F) in their earlier papers, but later the fraction of labeled analyte bound (b)] to analyte concentration ([H]), subsequently arguing that the response curve slope (dR/d[H]) is implicitly indicative of the detection limit of an assay (16). In contrast, Ekins et al. [e.g., (17)] explicitly defined the sensitivity of an assay as represented by the (im)precision of measurement of an analyte concentration of zero (i.e., \( \sigma_{b_0} = \sigma_{R_0}/(dR/d[H]_0) \), where R is the response variable (however expressed) and \( \sigma_{R_0} \) is the standard deviation of R and (dR/d[H]), is the response curve slope, both at zero dose), this quantity being the key determinant of the lower detection limit of any measuring system and thus of the actual ability of the system to determine, detect, or sense small amounts “of that which it is intended to measure” (18). The crucial difference between these two definitions is that the latter specifically requires detailed statistical analysis of assay data (to determine, either directly or indirectly, the magnitude of random errors in the measurement of the response variable, represented by \( \sigma_R \)). In contrast, the slope definition implicitly assumes constancy of such errors irrespective of any alteration in assay design (such as a change in the antibody concentration used in the system) or the manner in which the response curve is plotted.

The invalidity of this assumption, and thus the inadequate and misleading nature—indeed untenability—of the slope definition may be illustrated in many ways. For example, it may readily be shown (using either experimental data or the equations shown in Fig. 3) that plotting RIA data in terms of B/F implies that assay sensitivity (as thus defined) is increased by increasing the antibody concentration; however, the opposite conclusion is reached if the same data are plotted in terms of F/B. Plotting the curve in terms of \( b \) leads to a third conclusion, i.e., that maximal slope (sensitivity) is achieved using an antibody concentration binding 33% of a trace amount of labeled antigen (16) (this concentration being \( 0.5/K \), where K is the affinity constant governing the antibody/antigen reaction). In short, contradictory conclusions will be drawn regarding the conditions yielding maximal assay sensitivity in consequence of plotting identical experimental data in different coordinate frames (i.e., in terms of different dose or response variables) while disregarding random errors in the measurement of the
chosen response variable. [It should perhaps be emphasized that this phenomenon is not a unique feature of RIA but is a characteristic of all measuring systems. For example, which of two balances is deemed the more sensitive (slope definition) is likely to depend on whether the response of each balance to a weight placed on the pan is expressed in terms of the angular rotation of the beam or the movement of the pointer tip across the scale.] Misunderstanding of this issue contributed, for example, to initial uncertainties regarding the relative sensitivities of labeled antigen and the (radio)labeled antibody [immunoradiometric (IRMA)] assay methods discussed below. Although the latter were originally claimed (19) to be inherently more sensitive than RIA, the claim was based on a consideration of response curves per se and was not substantiated either theoretically or experimentally. Moreover, its validity was challenged by Rodbard and Weiss who, in 1973, published detailed theoretical studies (20) purportedly demonstrating that both labeled analyte and labeled antibody methods possess essentially equal sensitivities (albeit these authors failed to perceive the crucial distinction between competitive and noncompetitive assay designs—see below). Likewise, many observers were initially skeptical of the possibility of developing ultrasensitive microspot ligand assays (as described below in Principle Developments in the Ligand Assay Field) because the use of “vanishingly small” amounts of capture binding agent located on a small area contravenes widely accepted assay design rules stemming from the slope definition of sensitivity.

Although attracting less attention, disagreement also centered on the concept of precision adopted by Berson and Yalow (15, 16). This was similarly defined as the slope of the (b vs log[H]) response curve, without reference to the random errors incurred in the response measurement. Assays designed in conformity with this definition do not yield maximal precision according to the widely accepted scientific meaning of this term. In short, theoretical prediction or experimental observation of the effects of a change in assay components on the response curve slope in isolation (i.e., without consideration of concomitant effects on the errors incurred in the response measurement) provides a misleading indication of the effects of the change on assay sensitivity and precision and a fallacious basis for assay optimization and design. Nevertheless, an absence of appropriate statistical analysis of assay data has frequently characterized experimental studies in the ligand assay field, initially obscuring—among other consequences—the potential value of technical developments (such as the in vitro synthesis of monoclonal antibodies) or of new analytical concepts that have ultimately led to revolutionary improvements in assay design and performance.

Principle Developments in the Ligand Assay Field

Phase 1: Radioimmunoassay and Saturation Analysis

As indicated in Fig. 1, the publication of the insulin assay method of Yalow and Berson (1) essentially coincided with that of my method for serum thyroxine (T₄) (2). Dr. Yalow has revealed that the insulin assay emerged as a fortuitous by-product of studies on the immunogenicity of insulin. In contrast, the T₄ assay that I developed represented the outcome of an explicit quest for a general microanalytical method of sufficient sensitivity to assay substances (such as hormones) present at very low concentrations in body fluids. The reasons for, and a brief account of, this search may be of historical interest.

As a junior member of the Department of Physics Applied to Medicine at the Middlesex Hospital Medical School in London in the 1950s, I had the opportunity to carry out spare-time research relating to thyroid physiology. I also had the good fortune to have as a colleague and friend J.F. Tait, then collaborating with S.A. Simpson (later S.A. Tait) of the Courtauld Institute of Biochemistry on the studies that led to the discovery of aldosterone (21). Both were hampered by the lack of analytical methods of sufficient sensitivity to measure hormone concentrations in blood; this issue therefore constituted an occasional topic of discussion, although Tait and his co-workers succeeded in the early 1950s in applying the “double-isotope derivative analysis” technique to the assay of certain steroids (22), a method that—although cumbersome and time-consuming to perform—ultimately permitted many pioneering studies in the steroid field to be carried out.

Reflecting on this problem when walking home one evening in the spring of 1954, I perceived a possible general solution to it (Fig. 22) and a particular solution applicable to the assay of thyroxine based on the use (as a “saturable” binding agent) of the recently described specific thyroxine-binding globulin. Reaching home, I carried out a simple analysis of the implications of the mass action laws, deriving equations enabling the form of the dose–response curves in a system of this kind to be predicted (Fig. 3). Regrettably, these concepts were viewed with skepticism by my peers, and funding for their experimental implementation was denied. The consequent lack of suitable equipment or means to purchase and radiolabel the key reagents required led to the events that followed.

---

1 Berson and Yalow defined precision as db/(d[H]/[H]); this is equivalent to db/dlog([H]), the slope of the b vs log[H] response curve.
2 Although in recent published papers relating to the meaning of sensitivity, Dr. Pardue (11, 13) and I (10, 12) have disagreed on the definition and use of the word, we are essentially in agreement on the need for statistical analysis of assay data to assess the magnitude of statistical errors in the response variable, and hence the ability of an assay to determine a low analyte concentration.
3 Figs. 2–4 drawn in the period 1957–1960 and presented at a meeting of the Middlesex Hospital Medical School Research Society in London during this period.
Although work on the proposed methodology was, for the above reasons, precluded for the next 3 years, two fortuitous occurrences in 1957 enabled small-scale studies to commence without financial assistance. The first was the admittance to the Middlesex Hospital of a patient with thyroid carcinoma metastases who was treated with massive doses of $^{131}$I throughout the year (i.e., 200 mC at 2-month intervals), and whose blood radioactivity was being routinely monitored. In the course of these measurements, I unexpectedly observed that the radioactive material was largely protein bound, apparently comprising $^{131}$I-T$_4$. Having confirmed the latter chromatographically, I constructed simple electrophoretic equipment with which I was able to confirm the changes in distribution of labeled T$_4$ between specifically (i.e., thyroxine-binding globulin) bound and “free” (albumin-bound) moieties following the addition of varying amounts of exogenous, unlabeled T$_4$, thus enabling response curves relating the observed distribution to unlabeled T$_4$ concentration to be derived (Fig. 4).

The second fortuitous event of 1957 was a 6-week visit to my department by A. Harjanne, a clinical chemist from Helsinki, with whom I discussed the principles of the proposed analytical methodology and its requirement for high specific activity labeled hormone preparations. Dr.

---

**Fig. 2.** Labeled analyte binding assay, which I originally termed saturation analysis (later often referred to as competitive protein binding assay).

The specific saturable binding agent (Q) used in such a system may be an antibody, serum or receptor binding protein, enzyme, or oligonucleotide. It may be necessary to remove the analyte (P) from its biological milieu before reaction with Q. The distribution of labeled P between bound and free moieties depends on the concentration of P. (Slide prepared and presented at a meeting of The Middlesex Hospital Medical School Research Society, London, 1959).

---

**Fig. 3.** Equations (which I derived in 1954) governing saturation assay (including RIA) dose–response curves (17).

The concentrations of P (analyte) and Q (binding agent) are represented by p and q; $K$ is the equilibrium constant governing the binding reaction. Response curves may be plotted in terms of numerous alternative dose variables including (left) the free/bound ratio ($R_{FB}$) and (right) the bound/free fraction ($R_{BF}$). Likewise, various dose variables are commonly used—typically analyte concentration or log concentration. Values of $q$ and $K$ of 1 and 2 (arbitrary but consistent units, e.g., pg/mL or mL/pg) have been assumed in this figure. Reduction in the concentration (q) of binding agent (e.g., antibody) used yields a response curve of greater slope when curves are plotted in terms of $R_{FB}$, but of lower slope when plotted in terms of $R_{BF}$. Slide prepared and presented at a meeting of The Middlesex Hospital Medical School Research Society, London, 1959.
Harjanne suggested the use of the Conway microdiffusion flask (a device used by clinical chemists to determine iodine in blood) to radioiodinate substances such as thyroxine and proteins to high specific activities, notwithstanding the restricted facilities available. During the next 2 years, using either residual blood from samples occasionally derived from the patient referred to above or T4 labeled by the Conway diffusion method, I slowly obtained sufficient data to merit submission of a paper in 1959 describing the T4 assay. However, it was recommended by my departmental director that the method should not be presented as exemplifying a general analytical principle until it had been applied to another analyte; it was therefore not until 1962—having used the same approach to assay serum vitamin B12 (23)—that I published a description of “saturation analysis” as a general microanalytical technique (24).

The widespread adoption of this approach (sometimes termed “competitive protein-binding assay”, a term that clearly embraces RIA4) rapidly followed the publication of the studies by Yalow and Berson (1) and myself (2); however, it is unnecessary to elaborate further on these events, which are well known. The advantages and consequent popularity of the approach owed primarily to the specificity deriving from the “molecular recognition” features characterizing many binding reactions on which living systems rely, coupled with a high sensitivity stemming (in part) from the use of a high specific activity label (such as a radioisotope) to observe the reactions between small numbers of reactant molecules. However, the latter is not the sole factor contributing to the sensitivity of these methods; of equal importance is the choice of binding reagents with appropriate physicochemical properties and of optimal reagent concentrations. This choice is a key feature of ligand assay design, although—as indicated earlier—it emerged as a major topic of controversy arising from differing perceptions of the meanings of assay sensitivity and precision. Disagreement and debate on the meanings of these terms continues (10–13). Unfortunately, this cannot be brushed aside as a trivial semantic issue of no practical importance; many workers in the field continue to rely on the response curve slope alone as an indicator of the ability of an assay system to determine low analyte concentrations, and select assay reagents and protocols in accordance with this specious concept.

**PHASE 2A: IMMUNORADIOIMETRIC AND OTHER LABELED ANTIBODY ASSAYS**

Subsequent to the emergence of the first generation labeled analyte techniques described above, Wide et al. (3) in Sweden, followed shortly by Miles and Hales in the United Kingdom (4), developed in the late 1960s radiolabeled antibody methods (commonly referred to as immunoradiometric assays, or IRMAs), the studies by Miles and Hales being distinguished by their development of immunosorbent techniques for the preparation of relatively pure labeled antibodies. These methods constituted a reversion to the earlier “labeled reagent” techniques exploited by Avivi et al. (22), although they were distinguished therefrom by the use of reagents (i.e., antibodies) of much higher specificity than the labeled organic substances (e.g., acetic anhydride and, later, pipsyl chloride) used by Tait and his co-workers. The apparent motivation of Miles and Hales for adopting this approach was a belief that the labeled antibody approach would yield higher sensitivity (19), although they provided neither experimental nor rigorous theoretical evidence supportive of the claim, and as indicated previously, it was later disputed by Rodbard and Weiss (20).5

One disadvantage of the technique developed by Miles

---

4 As indicated in Fig. 2, saturation analysis was originally conceived as a general analytical approach, relying on any kind of specific binding agent (including antibodies). The term saturation analysis, although not ideal, was coined in an attempt to represent the generality of the analytical concept involved, avoiding any implied restriction either to the use of a particular class of binding reagent or to a type of label (although radioisotopic labels were, for many reasons, a natural choice). Antibodies had long been known (in the 1950s) to specifically bind antigens, and I regarded the possibility of their use in this context (as indicated in Fig. 2) as, therefore, entirely obvious. I was, of course, entirely unaware of the work of Yalow and Berson (1) in this area at the time Fig. 2 was first drawn and publicly presented.

5 IRMAs may—depending on their design and factors such as the nonspecific binding of labeled antibody—exhibit slightly greater sensitivity than correctly optimized RIAs; however, the improvement is marginal. On the other hand, the opposition by Rodbard et al. (20) to the claim by Miles and Hales (19) was incorrectly based on the supposition that the sensitivities of competitive and noncompetitive labeled antibody assay designs are identical.
and Hales is a concomitant loss in specificity because of the increase in relative potency of cross-reacting antigens as the antibody concentration used in the system is increased (25); however, this effect is largely annulled in assays of “sandwich”, or “two site”, design as devised by Yorde et al. (26), in which the analyte must be simultaneously recognized by two antibodies (directed against different epitopes on the analyte molecule) for an assay response to be generated.

Largely because of the absence of evidence demonstrating any improvement in performance of labeled antibody assays, coupled with their heavy consumption of antibody, assays of this design were almost totally disregarded by researchers and kit manufacturers for the next 10–15 years, albeit a commercial kit for the assay of hepatitis B antigen constituted a notable exception. However, the manufacturer’s adoption of the labeled antibody (sandwich) approach in this case appeared to be unrelated to sensitivity considerations but stemmed from the unusually large size of the antigen molecule, which caused difficulties in distinguishing between free and antibody-bound antigen moieties in a conventional RIA. Other exceptions included certain systems based on the use of enzyme labels (27), commonly referred to as enzyme-linked immunosorbent assay (ELISA); however, the primary motivation for their development likewise appears to have been unrelated to a desire to increase sensitivity (such methods commonly are relatively insensitive) but was to avoid the legal and practical impediments to the use of radioisotopically based methods in many laboratories. In short, although labeled antibody methods were first developed in 1967, the improved assay performance that, in certain circumstances, they may yield was largely unrecognized for many years, notwithstanding their occasional adoption for practical and logistic reasons.

PHASE 2B: ULTRASENSITIVE ASSAYS

The uncertainty regarding the relative sensitivities of conventional RIAs and IRMAs stemmed partly from the widespread misunderstanding of the meaning of sensitivity as indicated above (and hence of the experimental and statistical procedures required to determine the ability of a method to measure low analyte concentrations) and partly from a failure to recognize that the key distinction in this context is not between labeled antibody and labeled antibody methods per se but between competitive and noncompetitive assay designs. For example, certain labeled antibody methods can be portrayed as relying on the competition for a limited amount of labeled antibody between the analyte and exogenous unlabeled analyte coupled to a solid support, and can therefore also be described as competitive (28).

Much of the confusion relating to this issue can be dispelled by viewing the basis of ligand assays in terms of the “binding site occupancy” principle. When a “sensor” or “capture” antibody (or other binding agent) is introduced into an analyte-containing medium, sensor antibody binding sites are occupied by analyte molecules to a fractional extent that reflects both the equilibrium constant governing the binding reaction and the final unbound analyte concentration in the mixture. All immunoassays depend on measurement of the “fractional occupancy” of sensor binding sites by analyte, from which the analyte concentration in the system is deduced. However, occupied binding sites may be measured either “directly” or “indirectly” (i.e., by determination, in the latter case, of unoccupied sites). The statistical error incurred in the measurement of binding site occupancy (a key determinant of the sensitivity of an assay) depends on which of these alternative approaches is adopted, as does the optimal concentration of antibody used in the system.

In short, the crucial distinction between so-called competitive and noncompetitive methods does not center on which component of the system (if any) is labeled but stems from the particular approach adopted to determine binding site occupancy. For example, as shown in Fig. 5, in single-site labeled antibody assays of the type developed by Miles and Hales (4), the labeled antibody itself constitutes the sensor antibody, which after reaction with analyte may be separated into occupied and unoccupied fractions (using, for example, an immunosorbent comprising antigen or anti-idiotypic antibody linked to a solid support). If, after the separation of bound and free labeled antibody fractions, the signal emitted by labeled antibody bound to analyte (i.e., the “occupied” antibody fraction remaining in solution) is measured directly, the assay can be classed as noncompetitive. Conversely, if labeled antibody unbound to analyte (i.e., attached to the immunosorbent) is measured, then the assay can [as exemplified by Yorde et al. (28)] be described as competitive.

Competitive and noncompetitive assay methods differ in many of their performance characteristics (including their potential sensitivities) in consequence of the differing statistical errors incurred in the measurement of binding site occupancy. In general, direct measurement of occupied sites represents the better strategy, generally yielding higher sensitivity. (Analogously, it is preferable to determine a 1-cm length by measuring it directly rather than subtracting measurements of two greater lengths, e.g., 1 m and 99 cm, each of which is subject to error).

These simple concepts are supported by detailed theoretical analysis. Thus it can be shown (29) that, in a competitive assay design, the maximum sensitivity (i.e., the lowest attainable detection limit) that can achieved in a correctly optimized system is given by $e/K$ (where $e$ is
"Competitive" immunoassay "Non-competitive" immunoassay

![Diagram of immunoassays]

\[ Ab \rightarrow 0 \text{ for maximal sensitivity} \]
\[ (assuming \text{perfect separation of reaction products}) \]
\[ Ab \rightarrow \infty \text{ for maximal sensitivity} \]

Fig. 5. Binding site occupancy concept of immunoassay and other ligand assays.
All rely on measurement of binding site occupancy by analyte but differ in the way occupancy is determined. So-called competitive assays (left) rely on indirect measurement of occupancy by observation of unoccupied sites. Noncompetitive assays (right) rely on direct measurement of binding site occupancy. Note that labeled antibody assays may be either of competitive or noncompetitive design.

The relative "manipulation" error in the measurement of the assay response, and \( K \) is the affinity constant governing the antibody-antigen reaction), assuming that the label used is of infinite specific activity (i.e., that the statistical errors of the kind incurred in counting radioisotopic disintegrations are zero). In contrast, the corresponding expression relating to a noncompetitive assay is \( \epsilon/K \times b_n \) (where \( b_n \) is the fraction of labeled antibody that is nonspecifically bound) (7). Thus if \( b_n = 0.01 \), a noncompetitive assay is potentially 100-fold more sensitive than a competitive assay, other factors being equal. However, a different picture emerges when statistical counting errors are taken into account, as shown in Fig. 6 (7). For example, when a radioisotope such as \( ^{125}\text{I} \) is used to label the antibody, sensitivity is increasingly constrained (with an increase in the affinity constant of the labeled antibody) by counting errors consequent on the limited specific activity of the label. In other words, assuming that nonspecific binding of labeled antibody is kept to a very low level, counting errors are likely to replace manipulation errors as the principal limitation on assay sensitivity in a noncompetitive system.

These theoretical considerations indicate that the achievement of substantially higher sensitivities than are attainable by the "first generation" RIA techniques requires (a) use of a noncompetitive labeled antibody assay design, (b) use of a nonisotopic label of much higher specific activity than conventional radioisotopes, and (c) reduction of nonspecific binding of labeled antibody to the minimum practicable level (i.e., at least <1\% and ideally <0.01\%).

These conclusions led to my immediate recognition (9) of the potential importance (in an immunoassay context) of the Nobel Prize-winning development of in vitro techniques of monoclonal antibody production by Köhler and Milstein (8) and to the collaborative development by Ekins and co-workers (initiated ~1976) with Wallac Oy (then an instrument manufacturer) of the time-resolved...
fluorometric immunoassay methodology now known as DELFIA (5, 30). The collaboration was initiated, and its objective was defined, by myself (based on my then-unpublished theoretical analyses relating to assay sensitivity) and was actively continued until 1983–1984. Regrettably, no reference to myself or those of my colleagues participating in the collaboration was made in a number of papers published by members of the Wallac research team commencing in 1982 [e.g., (31–34)], nor (until 1984) was the “dissociation enhancement” approach to fluorescence measurement (on which the assays described in these papers relied and that had been developed by one of my colleagues, Dr. Salifu Dakubu, [see (35)]) disclosed.

Clinical Chemistry 44, No. 9, 1998

Fig. 6. Theoretically predicted sensitivities of competitive (a) and noncompetitive (b) immunoassay methods (represented by the SD of zero analyte measurements, expressed as molecules/mL) plotted as a function of antibody affinity ($K_{ab}$).

In the case of competitive assays, calculations are based on the assumption that the manipulation error (CV) incurred in the measurement of the assay response (e.g., fraction of labeled antigen bound) is 1%. The “potential sensitivity” curve assumes the use of an infinite specific activity label, i.e., that the statistical error in the label measurement is zero. The $^{125}$I-label curve indicates the sensitivity loss resulting from the counting of $^{125}$I for 2 min. With antibodies of an affinity $<10^{12}$ L/mol (the maximum generally achieved), little sensitivity is gained by using labels of higher specific activity than $^{125}$I. In the case of noncompetitive assays, potential sensitivity curves assume nonspecific binding of labeled antibody of 1% (upper curves) and 0.01% (lower curves). Corresponding $^{125}$I-label curves demonstrate the much greater sensitivity loss using a radioisotopic marker and the advantages of high specific activity nonisotopic labels in assays of such design (particularly if nonspecific binding is $\leq$0.1%). Arrows indicate sensitivities reported for noncompetitive immunoassays based on $^{125}$I (IRMA), and enzymes relying on fluorogenic (HS-ELISA) and radioactive (USERIA) substrates. [Reproduced from Ekins and Chu (7).]

clinical importance in the case of certain analytes, for example, thyrotropin, whose serum concentrations in thyrotoxic individuals not only fell below the limit of detection of the original RIA methods in common use in the early 1980s but were essentially indistinguishable from reference values. However, an equally important consequence of the development of ultrasensitive methodologies of this type has been a dramatic reduction in assay performance times, leading to the emergence of the fully automated immunoanalyzers that now dominate the immunodiagnostic field. Total incubation times in the order of minutes are typical, replacing the hours or days characterizing first generation competitive RIA methodologies.

To summarize this section, the use of nonisotopic labels of much higher specific activity than radioisotopes, when combined with noncompetitive strategies of binding site occupancy measurement, yield ligand assays of
greatly enhanced sensitivity and that require much shorter incubation times than competitive techniques relying on isotopic labels. The general recognition of the advantages of this approach (coupled with the increasing availability of monoclonal antibodies) has caused a major transformation of the immunodiagnostic field in the past decade and is the principal reason for a diminishing commercial interest in the original, isotope-based methods.

Against this historical and conceptual background, we may now turn to the development of ultrasensitive, miniaturized, multianalyte ligand assays based on microspot arrays.

**PHASE 3: MINIATURIZED, MICROARRAY-BASED, MULTIANALYTE ASSAYS**

The need to determine analyte panels in blood and other biological fluids has become increasingly evident in fields such as endocrinology, where resolution of certain diagnostic problems may require determination of the concentrations of several different hormones. However, this need is even more conspicuous in other areas, such as allergy testing, forensic medicine, the screening of transfusion blood for viral and other contaminants, and environmental monitoring. A complementary objective currently attracting much interest is assay miniaturization, not only because this permits measurement of many analytes in the same small sample but because of the concomitant reduction in assay performance times, the possibility it offers of “point-of-care” testing, and many other practical benefits.

Traditional methods of combining immunoassays within the same system have relied on the use of different labels permitting a number of individual binding reactions to be distinguished. However, the scope for extending the number of labels used in this way is limited. In contrast, spatial isolation of many immunoassays in the form of an array on a solid “probe” offers much greater flexibility, and—in macroscopic form—has been adopted as a format in certain commercial multianalyte assay kits. This approach also underlies miniaturized “microarray” technology.

The perception of the possibility of such technology followed from the theoretical prediction that highly sensitive immunoassays can be developed using far smaller amounts of antibody than have traditionally been perceived as necessary, implying that an individual assay can be performed using numbers of antibody molecules so low as to be confinable within a microspot a few square micrometers in area. This permits, in principle, the construction of exceedingly small arrays, each microspot in the microarray being directed against a different analyte.

As indicated earlier, all binding assays rely on the determination of binding site occupancy. Moreover, it may readily be shown (from the mass action laws) that the fractional occupancy \( F \) of a sensor antibody (or other specific binder) is given by:

\[
F = \frac{1}{K_{\text{Ab}}} + \frac{[\text{An}]}{[\text{Ab}]} + 1
\]

\[+ \frac{[\text{An}]}{[\text{Ab}]} = 0 \quad (1)\]

where \([\text{An}]\) is the analyte concentration, and \([\text{Ab}]\) is the sensor antibody concentration.

Furthermore, as \([\text{Ab}] \to 0:\)

\[
F \approx \frac{K[\text{An}]}{1 + K[\text{An}]} \quad (2)
\]

![Microspot ambient analyte immunoassay](https://example.com/microspot-assay.png)

*Fig. 7. Microspot ambient analyte immunoassay.*

After exposure to the analyte-containing medium of capture (or sensor) antibody located on a microspot, a second labeled antibody (Developing antibody) is used to determine sensor-antibody binding site occupancy, using a noncompetitive approach (left) or competitive approach (right).
Thus, $F$ generally varies with the amount of antibody present in the system; however, these equations also reveal that, when a sensor (or capture) antibody concentration of less than $\sim 0.01/K - 0.05/K$ (the antibody preferably being coupled to a solid support) is exposed to an analyte-containing medium, the resulting (fractional) occupancy of antibody binding sites reflects only the “ambient” analyte concentration in the medium and is independent of the total antibody present. It also follows that binding site fractional occupancy is independent of sample volume, provided this exceeds the volume required to ensure that $[\text{Ab}] < 0.01/K$. An immunoassay system conforming to these conditions may be termed an ambient analyte immunoassay (37).

A close similarity exists between these concepts and the operation of a simple thermometer. When placed in a liquid in a container, a small thermometer extracts heat from the surrounding liquid until it reaches thermal equilibrium, at which point the temperature of the liquid is indicated. Provided the thermal capacity of the thermometer is insignificant, the temperature it finally attains will be independent of its size. However, the introduction of a large thermometer (of high thermal capacity) into the liquid will cause a drop in the temperature of the liquid; thus, in these circumstances, the final temperature recorded will depend, inter alia, on the size of the thermometer and the volume of liquid to which it has been exposed.

Analogously, a small amount of antibody located on a microspot on a solid support [the total number of antibody sites being less than $v / K \times 10^{-8} \times N$, where $v$ is the sample volume (in milliliters) and $N$ is the Avogadro number ($6.023 \times 10^{23}$)] provides the basis of an ambient analyte immunoassay system sensing the analyte concentration to which it is exposed. For example, if $v = 1$ and $K = 10^{11}$ L/mol, the maximum number of binding sites causing negligible disturbance ($<1\%$) to the ambient analyte concentration is $6 \times 10^7$. Assuming antibody surface densities (when closely packed as a monolayer) in the order of $10^4$-$10^5$ molecules/$\mu$m$^2$, antibody microspots

Fig. 8. Antibody microarray. Each microspot is interrogated to determine its fractional occupancy by the analyte against which it is directed. A laser-based confocal microscope provides a sensitive method of determining fluorescent signals emitted from the spots. The ratio of $\alpha$ and $\beta$ fluorescent photons emitted from each spot reflects its fractional occupancy by the analyte against which it is directed and is solely dependent on the analyte concentration. The ratio is unaffected by the amount or distribution of antibody coated (as a monomolecular layer) on the probe surface. [Reproduced from Ekins and Chu (7).]
of areas in the order of 100-1000 \( \mu m^2 \) are sufficient to accommodate numbers of molecules falling within ambient analyte immunoassay limits.

After the exposure of a microspot to an analyte-containing sample, occupancy of antibody located within the spot may be determined by its exposure to a second (labeled) antibody reactive either with occupied or unoccupied sites, these representing noncompetitive and competitive strategies of binding-site occupancy measurement, respectively (Fig. 7).

Moreover, because the system depends on the determination of the fractional occupancy of sensor antibody binding sites within the microspot area, a useful strategy is to label the sensor antibody itself with a second label and to observe the ratio of signals emitted by the two labels. The two antibodies may be labeled with, for example, a pair of radioisotopic, enzyme, or chemiluminescent markers. However, fluorescent labels are especially useful because they enable arrays of microspots distributed on the surface of a “chip” or sample holder base (each microspot directed against a different analyte) to be optically scanned, thereby permitting multianalyte assays to be performed on the same sample. Thus, by scanning the array (using, e.g., a confocal microscope (Fig. 8) or charge coupled device camera) the fluorescent signals emitted from each spot can be determined, revealing the ambient concentration of analyte against which the antibody in the spot is directed. These principles underlie the Microspot\textsuperscript{TM} microarray technology currently under collaborative development by myself together with Boehringer Mannheim GmbH.

The proposition that microspot assays conducted in this fashion not only permit many different analytes to be simultaneously determined but may be more sensitive and rapid than conventional systems challenges accepted ideas in the field and needs brief explanation. Although it is often suggested (and appears intuitively obvious) that—to achieve the highest sensitivity—most if not all of the analyte present in a test sample should be captured by the capture (i.e., sensor) antibody (thus maximizing the signal generated from occupied sites), the increase in the background signal consequent on the use of an increased amount of antibody must also be considered. In short, increasing the area of a microspot (while maintaining the antibody surface density constant) decreases its fractional occupancy by analyte (albeit total occupancy, is increased) and increases the background signal generated from the spot. Thus the signal-to-background ratio decreases with increase in spot diameter, this phenomenon being likely to reduce sensitivity.

Clearly, if the microspot area were reduced to zero, no signal (other than instrument “noise”) would be observed, and the system would be totally insensitive. In practice, other statistical effects come into operation when the number of individual events (e.g., photons) observed by the detecting instrument is small, prohibiting the reduction of the sensor-antibody concentration below a certain lower limit. An additional constraint is the number of analyte molecules captured by the sensor antibody within the microspot area, which—if exceedingly small—will also cause statistical variations in the emitted signal. Thus miniaturization of the system beyond a certain limit reduces sensitivity. Nevertheless, using high specific activity labels, the optimal sensor-antibody concentration used in a noncompetitive assay design may be extremely low, although a reduction to below 0.01/\( K \) is likely to offer little or no benefit, assuming thermodynamic equilibrium is reached in the system.

Noncompetitive sandwich assay designs have generally been regarded as inapplicable to analytes of small molecular size (albeit certain workers, e.g., (38), have claimed to have overcome the difficulties involved), and the competitive approach may therefore often be obligatory. Although the sensitivity attainable is unlikely to be as high as in noncompetitive assays, theoretical analysis reveals that highest sensitivity is likewise achieved when

---

**Fig. 9.** Theoretically predicted noncompetitive microspot immunoassay sensitivities plotted as a function of the minimum developing antibody surface density (\( S_{\text{min}} \)) detectable within the microspot area. Values of capture antibody surface density (\( S \)) of 10\(^5\) binding sites/\( \mu m^2 \) and of developing antibody concentration of 1/\( K \) have been assumed. \( K \) and \( K^* \) are the equilibrium constants of capture (sensor) and developing antibodies, respectively. [Reproduced from Ekins and Chu (7).]

**Fig. 10.** Computer simulations of analyte binding to microspots (assuming typical analyte diffusion and antibody binding constants) show that equilibrium is reached more rapidly and that fractional occupancy of sensor antibody is at all times higher with smaller spots.
a very small amount of antibody located on a microspot is used (thus maximizing the fractional occupancy of antibody binding sites).

Detailed theoretical consideration of (noncompetitive) microspot immunoassay sensitivity (39) permits prediction of the sensitivities attainable when sensor antibodies of varying affinities are used as a function of $S_{\text{min}}$ (Fig. 9). The analysis indicates that the attainment of high microspot assay sensitivity requires a detector capable of determining very low surface densities of labeled developing antibodies, the close packing of sensor-antibody molecules within the microspot—by minimizing background—contributing to the fulfillment of this requirement. It also suggests that, using high specific activity labels, sensitivities of microspot assays are unlikely to be inferior and (depending on the measuring system) may be superior to the sensitivities achieved in macroscopic assays of conventional design. Researchers at Boehringer Mannheim have claimed that 0.01 labeled developing molecules/μm² have been detected in their experimental studies, suggesting that assay detection limits approximating $10^{-12}$ mol/L (i.e., $10^3$–$10^4$ molecules/mL) are attainable using the microspot approach.

These conclusions assume the establishment of thermodynamic equilibrium in the system; however, the velocities of binding reactions fall with a reduction of the concentrations of one or both of the reactants, and the time to reach equilibrium increases. Moreover, it is well known that diffusion constraints reduce reaction velocities when capture molecules are linked to a solid support. However, it is intuitively evident that the smaller the microspot, the lower these constraints. Theoretical analysis of the rate at which analyte molecules migrate toward, and bind to, an antibody microspot reveals that the (initial) antibody occupancy rate [per unit area of microspot (OR)] is given by:

$$\text{OR} = \frac{4r_m k_a [An] d_{Ab}}{(\pi r_m^2 k_a d_{Ab} + 4D r_m)}$$

where $k_a$ is the association rate constant (cm³·s⁻¹·molecule⁻¹), [An] is the ambient analyte concentration (molecules/mL), D is the diffusion coefficient (cm²/s), $r_m$ is the microspot radius, and $d_{Ab}$ is the antibody surface density (binding sites/cm²) (40). As $r_m$ tends to zero, the term $\pi r_m^2 k_a d_{Ab}$ becomes small compared to $4D r_m$, implying that OR approximates $k_a [An] d_{Ab}$. In other words, the rate at which sensor antibody molecules located within the microspot become occupied increases with reduction in $r_m$, ultimately approximating that seen in a homogeneous solution as $r_m$ approaches zero.

Detailed computer models reveal the full sequence of events following introduction of antibody microspots of different diameters into an analyte-containing solution; these models embrace inter alia the kinetics of the antibody-analyte reaction per se and the initial establishment

Fig. 11. Typical antibody (and oligonucleotide) microarray prepared by Boehringer Mannheim GmbH, using inkjet deposition technology.
of analyte concentration gradients in the solution. Such studies confirm that the smaller the microspot, the lower the diffusion constraints on the rate of analyte binding to antibody and the more closely the reaction kinetics approximate those characterizing a homogeneous liquid-phase reaction system (Fig. 10). These considerations reveal that higher signal-to-noise ratios are attained in a shorter time using a microspot format, implying that microspot assays are likely to prove at least as rapid as those of conventional macroscopic design.

Notwithstanding the validity of the above concepts, their exploitation on an industrial scale represents a formidable challenge. The industrial method of constructing microarrays developed by Boehringer Mannheim relies on small disposable polystyrene carriers onto which microspots are deposited using “inkjet” technology. Arrays of as many as 100–200 spots may be deposited in this manner on the flat bottom (~3-mm diameter) of the carrier wells, each droplet of reagent (antibody or antigen) solution having a volume of <1 nL and yielding a spot of diameter approximating 80 μm. Microspot array production by these means is presently effected in a fully automated unit at a speed of 10,000 carrier chips per hour. Spot size and array pattern and shape are controlled automatically by a digital imaging feedback device, allowing very precise microspot application (Fig. 11).

An alternative approach applicable to oligonucleotide and polypeptide microarray construction has been developed (originally for other purposes) by Fodor et al. (41), using combinatorial synthetic techniques. In this elegant approach, selected areas on a solid substrate are sequentially unmasked (by a photolithographic process) and exposed to a particular nucleoside (or amino acid in the case of polypeptide arrays). By these means, the set of all oligonucleotides of length n (numbering 4^n in total) are produced in 4n cycles. Among other advantages, this technique permits construction of large, high-density arrays, although the oligonucleotides present within individual microspots are likely to be of lower purity than when spots are formed by inkjet techniques using presynthesized material. Such in situ combinatorial techniques are, of course, not applicable to the construction of antibody arrays or to arrays of other molecular species not of simple single-stranded form.

Oligonucleotide arrays are clearly of potential importance for DNA analysis (42); however, such arrays can also be used as standard array templates, permitting individual researchers to construct antibody arrays of their own design, using antibodies to which complementary oligonucleotide sequences have been linked (Fig. 12).

Microarray technologies are still undergoing intensive development, and it would therefore be premature to

---

Fig. 12. An oligonucleotide microarray may be used to construct customized antibody arrays.
therefore, be little doubt that, if their commercial viability can be demonstrated, microarray technologies represent the next major advance in the diagnostic field. Clearly, however, this prediction will only be fulfilled if they match current methods in regard to sensitivity, precision, and speed of performance while yielding substantial improvements in convenience and cost.

Microarray-based DNA analysis technologies have attracted particular attention in the United States, their development being initiated by the Genosensor Consortium (established in 1992 as part of the US National Institute of Standards and Technology Advanced Technology Program). My own studies initially related principally to immunoarrays (in which area of application very high sensitivity is generally a prime requirement and whose achievement thus formed a key objective). Nevertheless, the future importance of “oligoarrays” for DNA analysis was evident from the outset, and the studies of myself and my collaborators have increasingly focused on this area of application.

Genetic analysis raises well-known ethical issues, which currently are the subject of much discussion and which are inappropriate to address here. It nevertheless appears likely that DNA analysis relying on oligonucleotide arrays to detect specific gene defects or genetically determined responses to particular drugs will feature prominently in the diagnostic armamentarium in the foreseeable future. Moreover, DNA analysis is of importance for a number of other reasons, for example, for tissue typing for transplantation surgery, “DNA fingerprinting” for forensic purposes, and the detection or identification of infectious agents (e.g., viruses) in human blood and tissues.

Meanwhile, antibody-based microarrays permitting the simultaneous determination of multiple analytes are likely to prove of particular value in certain obvious areas, e.g., in the identification of specific allergies and of viral antigens and antibodies in transfusion blood. Moreover, the requirement of the technology for small blood samples and the savings in time and cost of laboratory personnel consequent on the ability to simultaneously determine all clinically relevant biochemical indicators are likely to prove compellingly attractive to hospital laboratories throughout the world, currently under increasing budgetary pressures.

Summary and Conclusions

Ligand assays have made a major impact on medicine and related fields in the past 40 years. Initially time-consuming and laborious, the technology was transformed in the 1980s by the emergence of noncompetitive labeled antibody methods (relying on high specific activity nonisotopic labels), yielding higher sensitivities in much shorter times than the original techniques and opening the way to the development of the many automatic instruments that dominate the field at the present time.

For a number of reasons, international attention is increasingly focused on assay miniaturization, and large resources are currently being devoted to this topic in many countries and by many companies. Miniaturized microarray technologies of the kind described in this article are particularly suited to diagnostic applications because—although only requiring small sample volumes—they do not merely constitute miniaturized versions of conventional assay methods and thus do not themselves rely on the micropumps, microchannels, micrometer-sized reaction compartments, and other micromachined structures embodied in many of the microanalytical devices currently under development (although such devices may be necessary for sample preparation in certain analytical applications). Nor, in consequence, do they rely on the measurement of submicroliter (or nanoliter) sample volumes of such minuscule size that assay sensitivities are severely reduced and large statistical variations are likely to be encountered in the numbers of molecules (if any) that are present in the measured sample. There can, therefore, be little doubt that, if their commercial viability

References


