Searching for Molecular Solutions – Cited Notes

Chapter 8

These Files contain details on all references to this ftp site within **Chapter 8** of *Searching for Molecular Solutions*. The page numbers of the book where the reference is made are shown in the Table below, the corresponding page number for this file, and the title of each relevant section.

Contents:

Section 20: ***Maitotoxin Structure***

Cited on p. 278 of *Searching for Molecular Solutions*

This section provides a figure (Fig. 8. Na) showing the complex structure of maitotoxin, based on many polyether subunits.

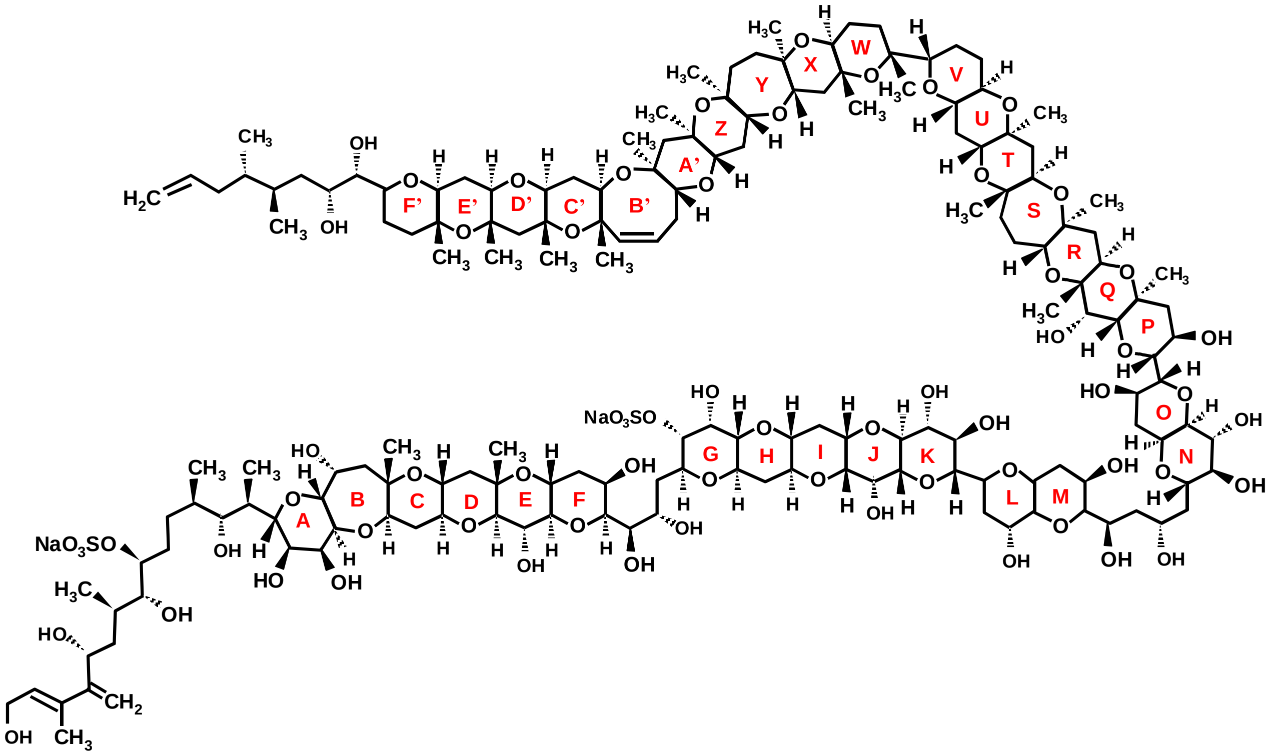


Fig. 8.Na

Structure of Maitotoxin[1-3]. The letter-code system for the polyether rings is as indicated in red.

Section 21: ***High-throughput Screening***

Cited on p. 287 of *Searching for Molecular Solutions*

High-throughput Screening Implementation and Methods

Even the chemical libraries of the most refined design will be practically ineffective if an efficient screening method is unavailable. And here the notion of performing screening in a rapid and systematic manner arises, almost always referred to as ‘high-throughput’ since its inception in the mid-1980s [4]. Though distinguishable technologies, chemical library design and high-throughput chemical library screening are therefore inherently intimately related [5,6]. High-throughput methods also feed back into the technologies for library syntheses themselves [7]. For these reasons, high-throughput screening is given a brief overview in this section, but high-throughput itself is simply an informational read-out concept which has very general applicability for molecular screening and diverse scientific enterprises. Automated high turnover screening processes can be designed for diverse libraries, which include collections of inorganic compounds and metal complexes [8,9], as well as more conventional libraries of organic molecules. Indeed, high-throughout approaches can be brought into play for screening for new drug targets themselves, which encompasses the area of whole-genome screens [♥].

This reminds us that screening for useful molecules by high-throughput methods is not necessarily restricted to single defined molecular targets, but can potentially employ any complex biosystem, up to the level of whole metazoan

[♥] An example of such ‘global’ processing is provided in the file SMS–CitedNotes-Ch9/ Section 29 (from the same ftp site) in the context of RNAi technology.

organisms. With conventional high-throughput screening where each library member is coded by its position in a spatial array (as In Fig. 1.2 of *Searching for Molecular Solutions*), each assay well can contain in principle a single target protein, whole cells, or entire small organisms, as long as a suitable read-out for functional activity of a library 'hit' is available [♥]. Trade-offs clearly exist here, though. A single molecular target may simplify screening assays but exclude detection of possible interfering effects occurring in more complex systems, while multicellular organisms can provide more realistic environments for the identification of candidate drug activity. High-throughput screening strategies for small molecule libraries have been established for use with the nematode *C. elegans* [10] and zebrafish (*Danio rario*) embryos [11]. On the other hand, as the screening system complexity grows, logistic restrictions can compromise the library size which can be realistically processed, especially if large and relatively slow-growing organisms are to be used. *In vivo* screening of mice for tissue-specific homing by small collections of compounds can be performed [12], but the screenable numbers are very low in comparison to fully *in vitro* conventional high-throughput approaches. The 'best' organism of all for screening the great majority of drugs is in fact *Homo sapiens* [13], but in this well-known species, controlled and systematic tests (commonly referred to as clinical trials) are not accurately deemed as 'high-throughput' procedures. Of course, drugs assessed in humans go through many levels of evaluation after initially emerging as leads, before entering even preliminary clinical testing [♣].

[♥] Another word with many meanings in different contexts, a 'hit' may be frowned upon in law enforcement circles, but definitely not in the very different world of high-throughput screening.

[♣] Yet still unpleasant surprises can occur, as seen in 2006 with the well-publicized severe problems in a Phase I trial of an antibody product (a CD28 [T cell coactivator] agonist) produced by the company TeGenero [14]. Clinical trials progress from Phase I through IV, with the numbers of patients involved increasing at each level. Adverse side-effects of some drugs have nontheless only become apparent after millions of clinical treatments, resulting in removal of such pre-approved drugs from the market [15].

These observations aside, most references to high-throughput screening have in mind high-level simultaneous (parallel) testing of large numbers of library candidates *in vitro*. Chemical library collections on the order of 10^7 compounds exist, with a million screenable by high-throughput processes in several weeks, subject to the type of assay required [♥] [16,17]. Though of respectable size, this volume is several orders of magnitude less than the number of library variants attainable with *in vitro* biological methods, such as mRNA display (Chapter 6 of *Searching for Molecular Solutions*), and (as noted in Chapter 8) is trivial in comparison to even conservative estimates of the size of small-molecule chemical space. Still, the objection of limited sampling of a huge potential total also applies with at least as much force for biological libraries of even small polypeptides, and the work striving towards maximization of chemical library diversity and chemical space coverage (Chapter 8) goes some way towards reducing this problem. While all modern library screening is dependent on relatively recent technological innovation, for high-throughput screening this is a compellingly direct feature. Screening at this level inherently requires automation in as many checkpoints as possible: if it can be carried out by human operators alone, it isn't high-throughput [♣].

[♥] High-throughput screening is also an expensive process, and requires chemical replenishment of the library stocks since (unlike biological libraries), the chemical library components are not directly replicable.

[♣] Some new technologies, especially in their early days, can require a tedious level of repetitive input by the operator, leading to a desire for either automation or extensive low-paid undergraduate student help. One such example is early DNA sequencing by the Maxam-Gilbert chemical degradation procedure, which prompted jokes to the effect that chimpanzees could be trained to perform its routine aspects and free up graduate students and postdocs for more creative pursuits. Unfortunately, such schemes never survive very long when reality checks are performed (if chimp intransigence and costs didn't kill such a project, animal rights activists probably would). But in the real world, DNA sequencing (using a number of different technologies) has long since entered the high-throughput era on genomic scales [18-20].

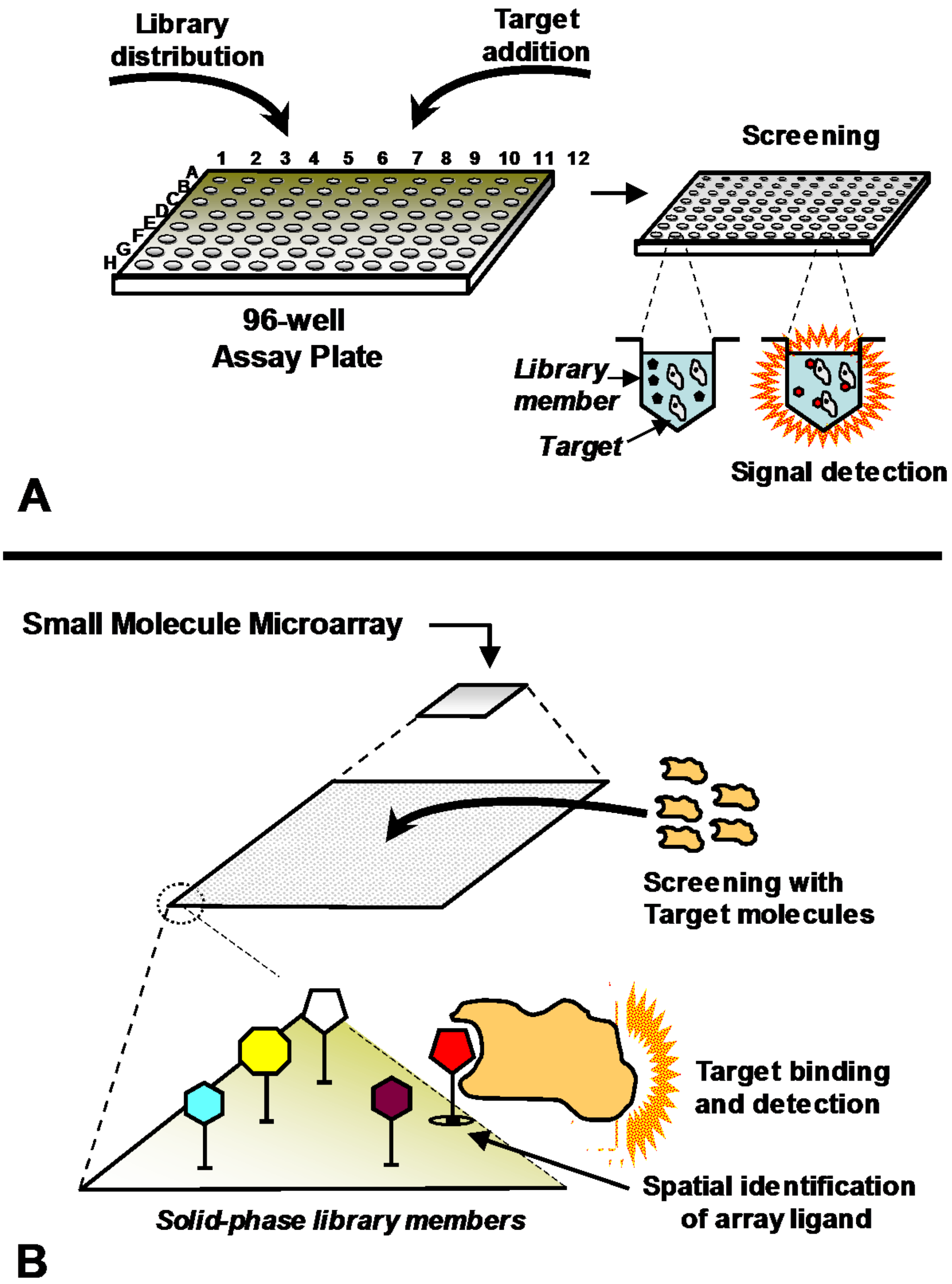

Library distribution
Target addition
1 2 3 4 5 6 7 8 9 10 11 12
A B C D E F G H
96-well Assay Plate
Screening
Library member
Target
Signal detection
A
Small Molecule Microarray
Screening with Target molecules
Target binding and detection
Spatial identification of array ligand
Solid-phase library members
B

Fig. 8.Nb

High throughput screening of small molecular libraries with microplate (**A**) and small molecule array (**B**) platforms. Each library member in (**B**) represents a microspot of predetermined quantities of specific compounds covalently bound to the solid-phase slide. Screening assays must result in a signal which assigns a position in the plate or microarray corresponding to a known library compound. A variety or enzymatic or fluorescence-based approaches can be used, such as a fluorescent tag on a protein target used in microarray analysis (**B**). Spatial encoding of the identity of a hit can be used in both cases. For example, in the microplate (**A**) the hit corresponds to plate coordinates H9, which in turn addresses the spatial encoding of the specific library compound within this well.

The required technical advances for the enablement of high-throughput screening can be broken down into three major areas: the physical arrangement of the target molecule or system to be used, the mechanism for the distribution of the library with respect to such targets, and the screening and hit-detection process itself. The first of these illustrates the interlinking of the library and the screening process. In conventional high-throughput screening *in vitro*, targets and individual library members are distributed into rectangular plastic 'microwell' plates with classically 96 (as depicted in Fig. 1.2 of *Searching for Molecular Solutions* and Fig. 8.Nb, but more recently 384 or 1536) physically separated compartment wells. In a standard arrangement, each well corresponds to the same target with a different library member, whose identity is spatially encoded by virtue of its specific position in the plate well grid. (In some circumstances, it is possible and desirable to screen pools of library members in a single batch, which clearly will dictate the physical requirements and well density required for the specific screening process in mind). The target molecule or system's environment is almost always aqueous, so one need for efficient high-throughput screening is automation of liquid handling for distribution to plate wells [16]. This, of course, is based on the assumption that target and ligands from within the library are both unconstrained within the 3-dimensional liquid volume within each microwell. With this kind of arrangement, complex targets such as living cells can be used, and a variety of functional screening assays for perturbation of cellular function can be run to search for drug-like activity.

Regardless of its size, the essential function of a microplate well is to act as a discrete compartment for the evaluation of a specific library member. Talk of ‘compartments’ might remind us of the development of the technology of emulsion-based *in vitro* compartmentalization, which was discussed in *Searching for Molecular Solutions* Chapter 4 in the same breath as display strategies. An emulsion droplet can act as a tiny ‘well’ for enabling high-throughput screening [21,22]. This is especially valuable for biologically encoded molecules, to preserve the link between genotype and phenotype which is the essence of display technologies, as we have previously observed. But unlike a rigid microplate, an fluid emulsion of microdroplets does not directly allow spatial encoding of the identity of its contents. It follows that emulsion compartmentalization is less useful for non-biological library screening unless it is accompanied by an encoding process through chemical tagging of some sort, a topic followed In Chapter 8 and extended within these Cited Notes for this chapter in sections below.

It is logical that increasing miniaturization of microwells should increase the cost-effectiveness of high-throughput screening. Moving much beyond plates with 1536 wells takes us into the realm of *microfluidics*, a burgeoning field given the strong incentives for miniaturization of a variety of biological assays. (Hence the catch phrase ‘lab on a chip’ [♥]). A microfluidic device contains channels, compartments and reservoirs for liquid distribution on a micrometer scale, usually in the form of microchips created by photolithographic processes [23]. As well as the opportunity for greatly increased parallel processing of samples, microfluidic

[♥] ’Lab on a chip’ has become a generic term for virtually any miniaturization of a chemical or biological process, not necessarily involving fluidic compartments (thus including sensor chips, etc.) [23].

compartments (with volumes potentially in the femtoliter range (10^{-15} liter, corresponding to a cube with sides of 1 micrometer [10^{-6} m]) require vastly lower transport times for mass (diffusion) and heat. Accordingly, regulation of compound concentrations and reaction temperatures can be achieved with great precision under such conditions [23]. Volumes in the picoliter (10^{-12} liter) range have been achieved for nucleic acid chemistries with microfluidic chips [24,25], and manipulation of even attoliter (10^{-18} liter) volumes with electrochemical 'syringes' appears feasible [26]. The rapid rate of progress in this field encourages the view that the need for macroscopic fluid handling by pipetting robots may be ultimately entirely superseded by microfluidic chips with integrated regulatable circuit channels [27]. Microfluidics can also assist the above-mentioned *in vitro* compartmentalization by greatly increasing the homogeneity of droplets in emulsions [23,28].

Despite the tremendous promise of microfluidics, for high-throughput screening of chemical libraries there is an alternative to grappling with tiny liquid volumes as screening wells shrink to the micro-level. Instead of solubilization in a liquid phase, library compounds can be rendered as solid-phase *microarrays.* Microarrays for simultaneous analysis of biological samples with a large number of separate nucleic acid or protein probes were noted within *Searching for Molecular Solutions* [♥]. In most of the latter cases, the range of targets and ligands are known; the specific information sought is the relative levels of a gamut of biomolecules in the original samples [♣]. In contrast, microarrays for high-throughput screening are designed by definition to provide new information about previously unknown target-ligand binding interactions. Since 1999, small

[♥] Also noted in the file SMS–CitedNoted-Ch6/Section 17; from the same ftp site.

[♣] For example, an array with oligonucleotides representing an organism's entire transcriptome can be hybridized with an RNA sample from a specific cellular source from the same organism, to examine relative genome-wide expression levels within the specific differentiated cell type of interest.

molecule microarrays [29] 'printed' onto slides [♥] have been developed and increasingly refined both in their preparation and screening methods [30,31]. The contrasts between high-throughput screening performed in microwells *vs*. microarray formats are depicted in Fig. 8.Nb. In contrast to microplates, tens of thousands of compounds can be printed onto a single microarray slide. As with microwell screening, individual library members embodied as a solid-phase array can be spatially encoded and identified (Fig. 8.Nb-B), but approaches coupling spatial positioning with chemical tag-encoding have also been developed, and these too are dealt with both in Chapter 8 and material within this Cited Notes section.

Although the small molecule microarray of Fig. 8.Nb-B depicts a protein probe, even whole cells can be used for the same target purposes [30,32]. Yet it might be noted that a solid-phase array format is restricted to binding-based detection, at least in the kind of arrangement depicted by Fig. 8.Nb-B. While microwell screening (Fig. 8.Nb-A) can use a wide variety of functional assays involving penetration of cells by active compounds, this is largely precluded when the library molecules are effectively tethered to the solid-phase surface. One way around this microarray limitation is to provide a means whereby members of the solid-phase library can be released under controlled conditions and then freely interact with cell targets in a region localized around the original microspot. Printing of small molecule libraries as conjugates with biodegradable polymers has been exploited to this effect [33].

The role of automation and robotic processing in high-throughput screening is emphasized by the above brief survey, and advances in these areas will continue to drive forward the field as a whole [34,35]. It may be a reasonably obvious point, but we should not forget that both the acquisition and analysis of library data are

[♥] One way to achieve this uses a chemically activated slide surface and a common reactive functional group for all library members, which enables each library compound to covalently bind to the slide when robotically microspotted onto a predetermined array position [29].

also heavily dependent on corresponding progress in information technology, or specifically ‘chemoinformatics’ in the case of chemical library deployment and the processing of screening data [36,37].

Screening and Intrepid Reporters

A word or two about procedures for the screening side of high-throughput would be worthwhile at this point. Again, the nature of an optimal screening assay is intimately associated with the nature of the target and the desired modulation of it, elicited by library hit compounds. The success of assays is dependent not only on their efficiency but also the spread of data obtained for both positive and negative controls (signal to noise), statistically quantifiable as their respective standard deviations. An assay giving ‘tight’ and reproducible signal to noise data may be preferable to another with greater variability, even if the former case produces signals whose average strength is lower. A statistical metric (Z’) for obtaining a rational measure of high-throughput screening assays based on signal and noise standard deviations has been developed and widely adopted for rating the likely productivity of new screening approaches [38,39]. Another important factor is the *dynamic range* over which an assay can be applied, which refers to the region in the ligand dose / assay response curve which maintains a well-defined functional relationship (usually linearity).

In order to accommodate the automation which is essential for high-throughput screening, the number of screening steps must be kept to an absolute minimum. An ideal is a one-step ‘homogeneous’ assay, requiring no other processing than addition of target sample and library members to each well of a screening plate [40]. This may contrasted with assays which require some kind of separation step, such as filter washing, to remove unbound probe. Treatment of a solid-phase array with a labeled target (as in Fig. 8.Nb-B) is a non-homogeneous procedure, since a washing step is necessary to remove unbound target. Dependent on the removal of unbound target and the identification of the sites of specifically bound target molecules, signal detection in such cases can be based on direct fluorescent labels or via an enzyme tag acting on substrates which in turn generate a measurable signal (Fig. 8.NcA).

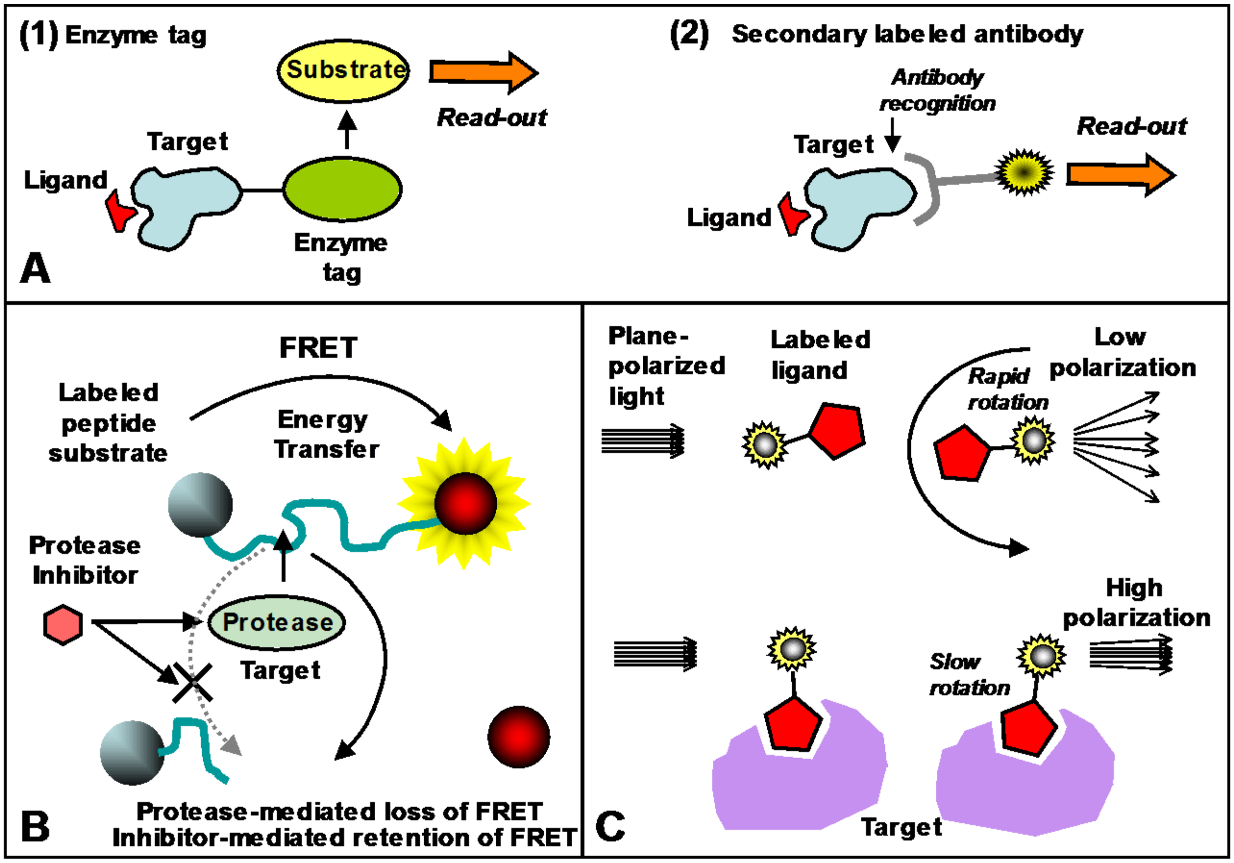


Fig. 8.Nc

Comparison of types of screens for high-throughput evaluation of libraries. **A**, Non-homogeneous screens requiring separation (washing or other treatment) steps, where the library elements are *immobilized* at spatially-defined sites. Protein targets directly conjugated with measurable labels (for example, fluorescent groups or radioisotopes), or conjugated with an enzyme (as shown) can provide a read-out with correct substrates (for example, those which generate a luminescent signal). The assayable tag can be indirectly provided through an antibody specific for the target (or some other target-specific binding agent). (**B** and **C**) Examples of fluorescent-based homogeneous screens. **B**, Protease-mediated loss of FRET signal from a labeled peptide substrate bearing a dual fluorescent label, as an assay for protease inhibitors. In the uncleaved state, an energy transfer between fluorescent donor and acceptors on the substrate results in measurable FRET fluorescence at a specific emission wavelength. Cleavage of the peptide separates the fluorophores and abolishes the FRET effect, such that the specific fluorescence is no longer emitted. Inhibition of the protease by a sought-after library member in turn results in restoration of the fluorescent signal. **C**, Use of fluorescence polarization measurements to assay for binding of labeled ligands to a protein target. In the lower section of this panel, the labeled ligand is depicted when complexed with a macromolecule (usually a protein) which strongly alters its rotational rate in solution, thereby also affecting fluorescence polarization, which can be read as a positive screening signal for a target hit.

In Chapter 6 Cited Notes (see the file SMS–CitedNotes-Ch6/Section 17; from the same ftp site), the versatility of fluorescence for a variety of applications in molecular detection was highlighted. Although the sensitivity of fluorescent measurement can be taken down to the level of single molecules [41-43], for high-throughput purposes, problems such as background autofluorescence and inherent fluorescence of library compounds must be taken into account. Certain sophisticated applications of fluorescence (beyond simple steady-state measurements) can cope with such problems in homogeneous screening assays [39,40,44]. Chapter 6 Cited Notes also described the phenomenon of Förster Resonance Energy Transfer (FRET), which can be applied towards high-throughput screening. The example of Fig. 8.Nc-B depicts the use of a special bilabeled peptide substrate, whose cleavage by a specific target proteases will block a FRET signal. Homogeneous screening of small molecule libraries for inhibitors of such a proteases will restore the signal, and serve to flag hits. In general, FRET can yield still more information if the emission decay of the fluorescence donor is analyzed [45]. This time-resolved FRET can distinguish different binding states and reduce fluorescent interference from library compounds themselves [39,44].

Another very useful fluorescent application is *fluorescence polarization*. When plane-polarized light is used to excite a small fluorescent molecule, its high rotational rate in solution will result in a low degree of polarization in its emission wavelength. But if the same fluorescent compound is bound by a much larger macromolecule, its rotation is slowed and its emission spectrum (in response to excitation with the same polarized light) retains significant polarization [44,46,47]. This effect can be applied for the detection of protein-ligand binding in homogenous assays (Fig. 8.Nc-C), and can also distinguish between fluorescence produced by proteins (such as green fluorescent protein) and small fluorescent molecules [46]. Fluorescence polarization has been also exploited in an interesting interface between the world of nucleic acid aptamers and small molecules, which is compatible with high throughput [48]. Since aptamers are usually relatively small molecules (in comparison with most proteins), they tend to show low fluorescence polarization when labeled with a fluorescent compound. A labeled aptamer pre-selected for binding a protein of interest will then show high polarization when bound to its protein target. Displacement of the aptamer by a small molecule then knocks down the polarization levels again, and affords a useful screening strategy for small molecules which bind the protein of interest at a similar site to the aptamer itself [48]. This then exploits the great power of aptamer functional selection for generating a foot-hold on small molecule screening, in effect ‘converting’ one form of molecular recognition into another.

If a small molecule is sought which perturbs a cellular process, it is often possible to design a screening assay whose signal is directly related to the process of interest, and which is obtainable via a cell-based assay. ‘Reporter’ systems for high-throughput screening can co-opt a wide range of cellular processes for serving as indicators for the desired functional activity, and can be implemented at the level of transgenic organisms such as zebrafish [11]. But screening must be carefully tailored to the project aims, illustrated by an example of one type of cellular reporter assay (depicted in Fig. 8.Nd) for a screen searching for small molecules modulating the expression of a protein of interest. In this designed system, a suitable cell line has been stably transfected with a construct encoding the protein of interest under the control of its normal promoter region. The reporter signal derives from fusion of the protein of interest with a small inherently fluorescent polypeptide (such as green fluorescent protein).

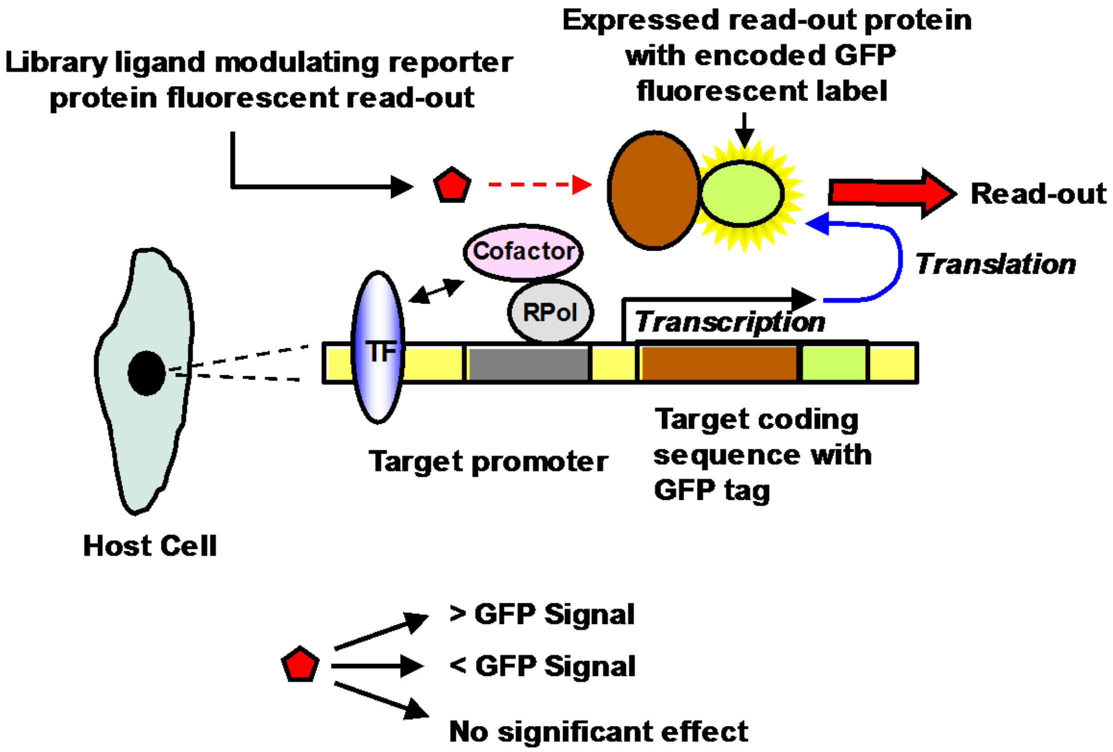

Library ligand modulating reporter protein fluorescent read-out
Expressed read-out protein with encoded GFP fluorescent label
Read-out
Cofactor
RPol
TF
Translation
Transcription
Target promoter
Target coding sequence with GFP tag
Host Cell
> GFP Signal
< GFP Signal
No significant effect

Fig. 8.Nd

Example of a homogenous cellular-based assay for small molecule modulation of expression, where the assay read-out is fluorescence from the green fluorescent protein (GFP) tag fused to a target of interest. Note that a library ligand in principle can increase the read-out in numerous ways, including increased transcription, mRNA stabilization, enhanced translation or protein stabilization. A positive signal as a primary hit then requires considerable additional work to investigate the level at which the effect is operative. TF = transcription factor; Rpol = RNA polymerase.

If a library molecule activated transcription in some way, then expression of the tagged protein should be augmented with an accompanying increased signal from the associated fluorescent tag. But it is important to note that a positive read-out at the protein level in this particular example does not at all guarantee that the responsible compound is a transcriptional activator, and might make us recall the 'First Law of Directed Evolution' (noted in Chapter 4 of *Searching for Molecular Solutions*) 'you get what you screen for'. An enhanced fluorescent signal correlates with increased steady-state levels of tagged protein, but as well as transcriptional modulation, this could result from mRNA stabilization, translational effects, or direct stabilization of the protein. The point is that primary hits arising from the use of such reporters require additional evaluation to confirm the level at which they are acting. Of course, the underlying goal in such a screening project is very significant in determining the potential value of candidate compounds. For example, if one is searching for specific stabilizers of mRNA, only a small subset of the hits from a screen as in Fig. 8.Nd are likely to be useful for follow-up, and the screen itself is probably a poor choice for the intended target. On the other hand, if all one seeks is an agent increasing steady-state protein expression levels by any pathway, then many true-signal hits should prove worthy of further investigation. Screens for small molecules enhancing protein expression levels in cultured mammalian cells have implicated both transcriptional activation and other mechanisms as responsible for the positive effects of specific compounds [49].

An extended discussion of all screening options available for high-throughput processing of libraries is well beyond our scope, but suffice it to say that this field is large and expanding [16,39,50,51]. As well as fluorescent-based technologies, advances in nuclear magnetic resonance and mass spectroscopy have become significant options [16,52]. But however high-throughput screening is undertaken, it would seem logical that the more information regarding compound performance and effects which can be obtained, the tighter the focus on contenders with the highest potential for passing through into the lead stage. Automated imaging techniques are well-placed for the maximization of assay informational content from the screening of small molecules. The intracellular behavior of multiple proteins (or other macromolecules) with separate fluorescent tags can be simultaneously monitored by fluorescence microscopy, with quantitation possible by image analysis software. (Usually the *in situ* proteins of interest are visualized by post-staining with specific antibodies themselves bearing distinguishing fluorescent markers). When such multiplex analyses are done on a high-throughput basis, the general approach (including associated instrumentation, reagents and software) is termed *high content screening* [53-56].

For drug development, 'additional information' of special relevance conforms to the ADMET criteria noted in Chapter 8 (absorption, distribution, metabolism, excretion and toxicity). High-throughput screening with large compound libraries may give sizable numbers of primary hits, a great many of which will prove to have no value for further development. Many of these 'non-starters' fail to meet one or more of the ADMET factors, and for this reason the integration of ADMET assessment with high-throughput screening at as early a stage as possible is highly desirable [57,58]. Properties such as solubility, hydrophobicity, cell permeability, stability and others can be measured with assays set up in parallel with the primary biological screen [57]. High-throughput evaluation of specific compounds for their ADMET performances has been addressed by automation of chromatographic and spectroscopic techniques [4,57]. Another important practical factor to take into account is *in vivo* drug-drug interactions [♥]. In a high proportion of drugs, a group of proteins termed cytochromes P450 are

[♥] For example, despite the value of the drug cimetidine (Fig. 9.2 of *Searching for Molecular Solutions*), it lost ground to competing drugs through its propensity to promote deleterious effects with co-treatments of certain drugs [59,60].

responsible for their metabolic processing, and partial inhibition of P450 activity by one drug can adversely affect the activity of another co-administered drug. Consequently, high-throughput screening for candidate drug effects on P450 function is another early stage step increasingly instituted [60,61].

There are many logistic issues with high-throughput screening of small molecule libraries which extend further than the range of this brief overview. We can note in passing some basic variables which must be contended with, including compound concentrations, choice of initial compound solvent, and number of replicates [4]. With respect to the first of these, a direct relationship has been observed between the number of primary screening hits obtained and increasing compound screening concentrations (up to a point). Higher screening concentrations can increase the rate of false-positives but also potentially facilitate the identification of novel classes of functional molecules [39]. One way to considerably extend the screenable primary library size in a high-throughput study is to use *compound pools*. Here, instead of one compound per well (or equivalent), groups of compounds are screened together, and wells showing hits are then progressively subdivided into smaller sets until individual active constituent molecules are identified [♥]. This process is possible where the composition of each pool is precisely defined, enabling the compounds which compose an active pool to be placed into smaller pools as desired. Compound pooling may not be strictly necessary for screening a given chemical library if high-throughput technology can cope with rapid library evaluation on a single-compound basis, but pooling may become a desirable option through reducing the initial screening costs, which are often very significant [62]. But on the other hand, this approach needs to be taken with caution to avoid introduction of artefacts. In principle, in a complex target system the action of one compound might antagonize the effects of another member within the same pool, or pool

[♥] This is directly analogous to a cloning procedure in molecular biology termed sib-selection (and referred to in Chapter 1), where pools of clones are screened for a function, and successively split into smaller and smaller sub-pools until unique clones are identified.

members might chemically interact with one another. With respect to the latter, rational pool design steps can be taken in order to minimize the problem [62].

While the first step in performing high-throughput screening is the identification of hits, it should be kept in mind that a hit is simply a positive primary screening signal, which needs to pass a subsequent series of evaluations, the ‘hit-to-lead’ process [63]. After confirmation as a true hit (eliminating false positives), continuation down the hit-to-lead pathway depends on its possession of favorable drug-like features and passing ADMET criteria as noted earlier. (This may involve rational side-group modifications aimed at improvement of specific criteria found wanting in an otherwise promising candidate compound). A designated lead molecule is then subjected to chemical structure / functional ‘evolutionary’ optimization, which may involve using fragments of the original molecule as the basis of further derivatization, subject to its initial size and complexity [63]. Extensive modifications of the original structure may be required. In contrast to leads resulting from artificial combinatorial chemical libraries, it may be noted that many natural products have passed from the lead stage to marketed drug without any alteration [64], again most likely as a result of the evolutionary ‘pre-screening’ of natural compounds for compatibility with protein folds.

In Chapter 8 it was noted that initial enthusiasm for combinatorial chemical libraries somewhat oversold their potential at first, but their value has become increasingly appreciated as the associated technologies have matured. As would be expected from the intertwining of such libraries and the means for their practical evaluation, the same caveat has been observed with the evolution of high-throughput screening technology itself [65]. Both chemical libraries and their screening by high-throughput strategies have nevertheless become entrenched as useful options in many fields of molecular discovery.

Gunning for Tumors the High-throughput Way

By considering the range of applications for chemical libraries and their screening by high-throughput methods, we are in the end returned to the questions of target druggability. It was also noted in Chapter 8 that at least a subset of cases previously regarded as intractable drug targets (especially involving protein-protein interactions with large contact surfaces) have yielded to concerted efforts to find small molecule inhibitors. It might seem that one could invent an aphorism ‘seek hard enough and ye may find a small molecule for the task at hand’, which might in principle have some validity but falls down in the face of the vastness of chemical space. As we have also seen, the success of an empirical screen is greatly assisted by rational implementation steps during its early stages. But the ultimate utility of a drug *in vivo* depends on target choice, its relevance to the underlying disease state, and its safety. Nowhere else are these dilemmas more acute than in the search for anti-cancer drugs, whose track record is much poorer than other therapeutic areas [66]. A majority of ‘oncodrugs’ with known efficacy in cancer also stand out through their failure to satisfy the usual criteria (such as the Rule-of-Five) for orally bioavailable drugs [67].

The problem inherent with cancer is defining targets unique to the tumor cells, or at least targets whose inhibition will cause the lowest possible toxicities to normal host systems [66]. Screening of chemical libraries for agents perturbing cell division processes can yield useful biological probe molecules [68,69], but for a compound to be therapeutically beneficial against cancer, clearly tumor selectivity is required. This point is amplified precisely by rare successes where a tumor-specific target has indeed been identified, leading to successful drug development (such as the kinase inhibitor Gleevec [♥]). Given the importance of kinases in intracellular signaling for control of growth-related processes, and the demonstrated druggability of these enzymes, both rational design and high-

[♥] Gleevec is a paradigm for such success, which is considered in the file –Extras-Ch9/Section A14; from the same ftp site.

throughput screening approaches have been brought to bear in an effort to derive specific cancer-relevant kinase inhibitors. A number of homogeneous assays compatible with high-throughput have accordingly been adapted and applied towards kinase inhibitor screening [70]. High-throughput searches for inhibitors of the Raf kinase (a component of an important signal transduction pathway) led to the development and approval of the inhibitor sorafenib for therapy of renal tumors, with the likelihood that it may prove beneficial in other cancers as well [71].

Since the aim of cancer treatment is to selectively kill cancer cells, it should not be surprising to hear that many workers have aspired to tip transformed abnormal cells down the pathway of programmed cell death. Close on the heels of the insight that most (if not all) eukaryotic cells have in-built mechanisms for self-destruction (apoptosis) came findings that many tumors subvert such apoptotic processes in order to survive. Over-riding tumor anti-apoptosis signals might then provide an avenue towards tumor annihilation. Out of a vast literature on this subject, for the present purposes we can note that chemical library screening has been applied towards the isolation of candidate apoptosis-inducing compounds potentially suitable for anti-tumor therapy [72,73]. But another reason for choosing apoptosis as an example is the stark conundrum that one clearly cannot simply activate ‘programmed’ cell death indiscriminately, as prompting normal cells to follow suit along with their transformed counterparts will hardly produce an optimal therapeutic result. Thus we are again returned to the core difficulty of finding a truly tumor-specific treatment.

A useful approach to this problem has been application of the principle of ‘synthetic lethality’, which is defined as occurring when altered expression of either of two genes is not lethal, but simultaneous change in both such genes is incompatible with survival [74,75]. Either loss of a gene, or its aberrant overexpression, can thus render a cell vulnerable to a change in another gene [♥]. Pairs of cells with and without such genetic changes are then screened in tandem to identify agents which differentially affect cell survival (as depicted in Fig. 8.Ne below). If the cells are genetically identical except for a specific tumor marker alteration and provided with distinguishing fluorescent labels, then the screening system is at its most efficient for identification of genes whose alteration confers a synthetic lethal phenotype [77]. Practical application of this screening philosophy has yielded promising candidates for renal cell tumors [78], and has considerable general promise. Returning to apoptosis in this context, numerous targets controlling apoptosis have been defined, [79], and some apoptotic gene product / drug interactions are interpretable along the lines of the synthetic lethality paradigm. In the end, tumor genomic instability and heterogeneity drives a Darwinian selection for successful variants [80], and this may prove a limitation for the penetrance of drugs obtained by synthetic lethal screening [66].

Much more could be said in this area, but let's take a different tack and consider that proteins, or protein-based systems, are most often the screening targets for modulation by small molecules. Even where the screen is at a whole-cell phenotypic level where the molecular target is not initially defined, proteins as ultimate targets are commonly an expected outcome. But we should not forget the relevance of other biomolecules as well, and a very active field of research is the examination of specific RNA motifs as useful sites for drug molecule binding. Although we have seen that protein druggability is likely to be more wide-ranging than once believed, a large section of the proteome remains difficult to modulate with small molecules, if not intractable. From this observation, a very salient point is that RNA molecules should take a higher profile as potential drug targets [81].

[♥] These kinds of interactions have been analyzed through global mutational anayses in yeast [76], as touched upon in *Searching for Molecular Solutions*.

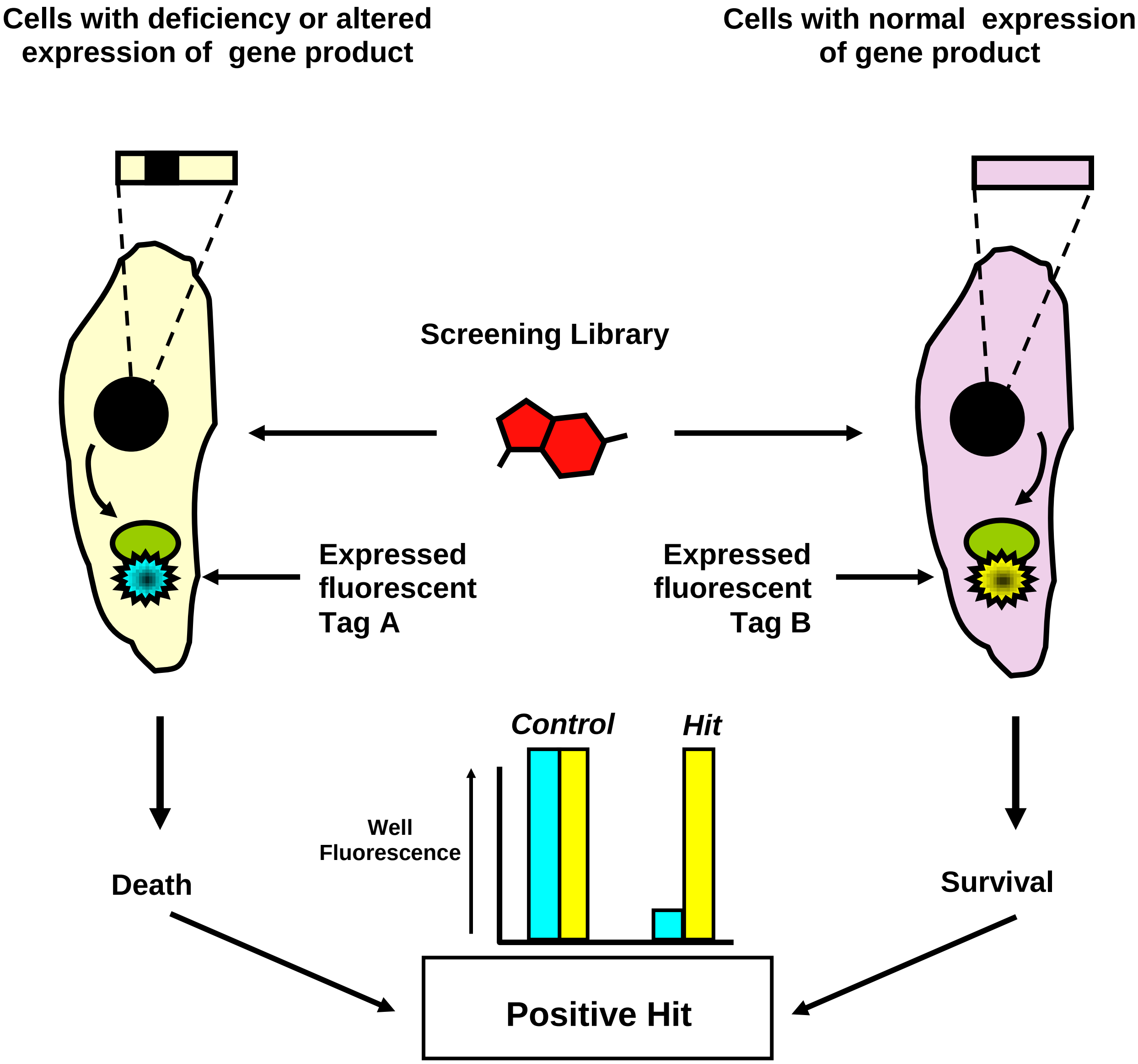

Cells with deficiency or altered expression of gene product
Cells with normal expression of gene product
Screening Library
Expressed fluorescent Tag A
Expressed fluorescent Tag B
Control
Hit
Well Fluorescence
Death
Survival
Positive Hit

Fig. 8.Ne

Screening for anti-cancer compounds with cells isogenic except for aberrant expression of a tumor marker in one case, and identifying fluorescent tags. Only library members which are differentially cytotoxic to the cell with aberrant marker expression are scored as hits (indicated by bar graph schematic of fluorescence in a microwell giving a positive signal). The target of a true-signal hit is thus a gene product whose expression is dispensable in the normal cell but essential in the aberrant cell owing to additional genetic alteration.

The druggability of RNA *per se* is not controversial, since a number of antibiotics exert their function via binding to prokaryotic ribosomal RNAs [82]. Yet targeting structural motifs in other RNAs remains an ongoing challenge, which (as in so many cases) can be met by a combination of rational and empirical high-throughput screening approaches [82]. Where do tumors fit in with this? Firstly, if the ‘holy grail’ of producing small molecule-based binding and inhibition of specific mRNAs could be routinely realized [♥], then tumor markers could be targeted, along with many other applications. Also, gene regulation through RNA mechanisms (micro-RNAs [miRNAs] in particular) are increasingly recognized for their general importance in health and disease, and control of miRNAs by specific small molecules would find an equally broad range of utility, certainly not excluding tumor cell targets.

Although much more could be said about the successes of high-throughput screening of chemical libraries, we should also think about the limitations of this branch of molecular discovery as well….

High-throughput Library Blues

Once again we lump small molecule libraries and the means for their rapid screening as a ‘package deal’, and problems with high-throughput screening can arise from either the design of the library or the design of the screening process. With respect to the latter, we have already considered that the desired end-point

[♥] In this context we should note other approaches for specific blocking of gene expression (especially RNAi), which is briefly noted in Chapter 9. But if a safe drug-like small molecule had at least equal value for expression control as a nucleic acid-based alternative, the small molecule would probably win in the marketplace.

should determine the nature of the screen as much as possible (as in the example of Fig. 8.Nd). Considering the costs of a large-scale high-throughput screening project, careful planning of the best screening approach is worth the time investment. (As previously noted, replenishment expenses are a very significant issue with large chemical collections of any sort [17]).

One way to look at the maximization of high-throughput screening potential is to consider the entire global repertoire of small molecular entities, obviously a constantly changing variable, but estimated as $>10^8$ synthesized compounds [83]. But of course not all of these are available to any one party, even the largest players in the pharmaceutical industry. Even if 50% were obtainable by an ambitious group, costs of ‘ultra-high-throughput screening’ become very significant, not just in billion-dollar terms but also from the viewpoint of such important issues as computational analyses and data storage space [83]. If the final success rate through successive rungs in a checkpoint ladder (broadly including screening library / primary hits / confirmed hits / leads / drug candidates, proceeding ultimately to new marketed drugs) is as low as 10^{-6} (as a fraction of the initial screening effort), then the end-products would need ‘ultra-blockbuster’ status in order to justify global-repertoire mass screening. Very high attrition rates may be acceptable in any screening process if a high proportion of a total population can be evaluated, yet 10^8 compounds, as we have seen, is much less than a drop in the bucket in terms of the size of chemical space….

Again these considerations suggest that brute-force screening alone is not the best answer, and return us to the arena of ‘smart’ library design and the computationally-assisted maximization of chemical library diversity. Library design can also attempt to focus on the sometimes elusive quality of ‘lead-likeness’, and accordingly eliminate ‘non-leadlike’ compounds through the filtering of specific functional groups [84,85]. A crucial issue is that irrelevant compounds do not merely take up library space which could be better deployed, but such ‘chaff’ molecules can also contribute a significant portion of false-positive signals during primary screening, or mis-hits (sometimes rendered without the hyphen as ‘mishits’, a word which for some reason tends to convey accurate negative connotations within itself). And false-positive ‘nuisance’ compounds are an expensive distraction [85].

Defining the origins of artefactual screening signals, and the nature of the compounds which generate them, has clear practical benefits and scientific interest as well [86]. In screening for enzyme inhibitors, multiple mechanisms can result in apparent positive signals which have no relevant drug validity. Compounds which directly interact with the components of a screening assay are one source of such problems, and in this case at least the formal possibility exists that an alternative assay will be capable of accommodating them. More intrinsic problems exist where compounds are reactive with enzyme functional groups, or have ‘privileged’ structural qualities which enable them to cross-bind to multiple members of a protein family [87]. Some structurally-unrelated molecules can act as non-specific promiscuous enzyme inhibitors through the formation of colloidal aggregates, which can be reduced by high non-specific protein concentrations or detergents [87-89]. A particular cautionary note in this regard comes from observations that at high concentrations, some drug-like molecules and even certain known useful drugs can form promiscuous aggregates [86]. Such information is clearly relevant to the design of screening conditions and library drug assay concentrations. And although the aggregation phenomenon leads to non-specific enzyme inhibition *in vitro*, it has been suggested that aggregation may assist the *in vivo* oral bioavailability of certain useful drugs [90]. Drug aggregation and associated colloidal inhibition mechanisms may also prove useful in blocking pathological protein filament formation, as with amyloid fibers associated with Alzheimer’s disease [91].

A related issue to non-specific drug activities is that of ‘off-target’ effects, or *in vivo* drug specificity, which is described further in ‘The Interactome and Biological Parsimony’) (see the file SMS–Extras-Ch9/Section A15, from the same ftp site).

Section 22: ***Genomics and Chemogenomics***

Cited on p. 290 of *Searching for Molecular Solutions*

Genomic Introduction – From Stone Age to 'Ome Age

We live in a very narrow and unique slice of history that might, with considerable accuracy, be called the Age of Genomes [♥]. The justification for this statement lies in the stunning rate of genomic sequence acquisition in recent times, mostly very recent indeed, but beginning about three decades ago (Fig. 8.Nf). When a genome has been accurately sequenced, confirmed and annotated, the information is there for a huge range of applications. As with climbing a mountain [♣], once a genome is 'conquered' by sequencing, the achievement has been accomplished and cannot be replayed again as a 'first'. While the number of distinct genomes within the biosphere is huge (and mostly prokaryotic), already many of the major genomes of interest to humans have been sequenced (some of which are shown in Fig. 8.Nf; but numerous others are not included). Certainly the acquisition of genomic data for a wide variety of species will be an on-going enterprise into the future, but within a sliver of time less than half an average human life-span, genomes of many of the most scientifically and economically important organisms have yielded their sequences, and this will never be repeated in human history. Be glad you are here to witness it.

[♥] We can define a genome as the complete nucleic acid content (usually DNA, but not always) of an organism which specifies its structure, growth and development. An exception exists within most eukaryotic cells for organelles such as chloroplasts and mitochondria which have their own self- replicating genomes. Also, independently-replicating episomal entities such as plasmids (in any type of cell) can be spoken of as having their own (usually small) genomes.

[♣] The mountain-climbing analogy also seems apt since the tallest mountain on Earth, Mt. Everest, was scaled by Hillary and Norgay in 1953, the same year that Watson and Crick published their famous paper on the structure of DNA. Note also that only a 25-year gap exists between defining the structure of DNA itself and the first DNA viral sequence determinations (Fig. 8.Nf; in 1976 the RNA bacteriophage MS2 was the first phage genome to be sequenced).

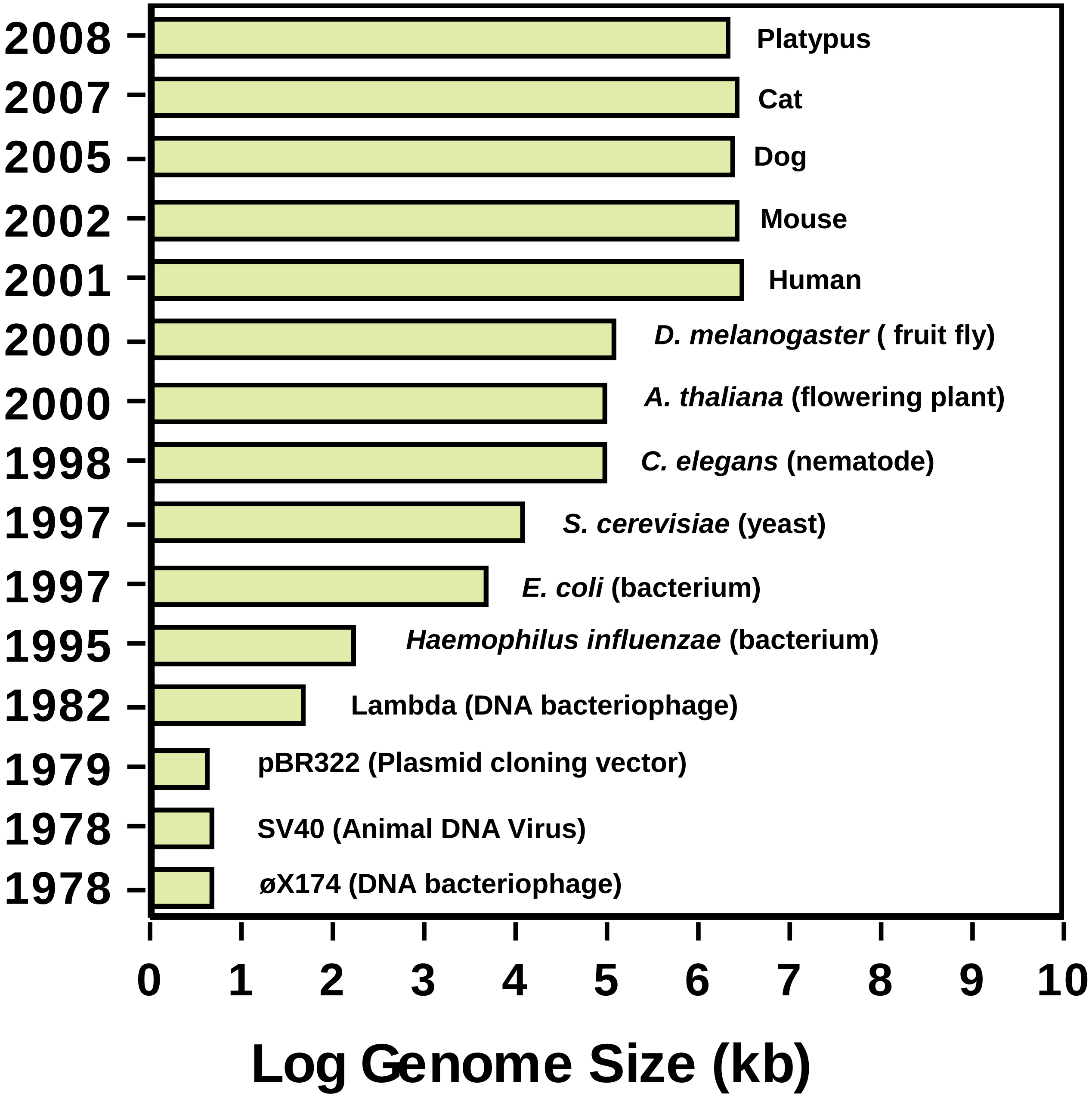


Fig. 8.Nf

Some milestones in genomic sequencing over a thirty-year period, with specific years of completion of sequence projects shown on the Y-axis. The genomes are scaled as the logarithms of their sizes in kilobases (kb). Additional points of interest and background information:

øX174: The sequence of this phage [92] revealed the first instances of genes with overlapping reading frames. The **SV40** virus encodes within its sequence [93] the Large T antigen (considered in more detail in the file SMS–CitedNotes-Ch9/Section 30; from the same ftp site). The plasmid **pBR322** is a historically important cloning vector derived from the *E. coli* plasmid ColE1. Although little used in its original form today, parts of its sequence [94] are widely distributed in more recent vectors. The phage **lambda** has provided a wealth of information on gene regulation, as well as serving as a cloning and display vector. Its sequence determination [95] was a landmark event at that time. The genome of ***Haemophilus influenzae*** was the first from a free-living organism to be sequenced [96]. The sequence of ***Escherichia coli*** [97] was of great significance given its central importance as a laboratory tool. The first eukaryotic genome sequence determination came from the important yeast model organism ***Saccharomyces cerevisiae*** [98]. Moving on to multicellular eukaryotes, the genomes of the experimental organisms ***Caenorhabditis elegans*** [99] (nematode worm), ***Arabidopsis thaliana*** [100] (flowering plant) and ***Drosophila melanogaster*** [101] (fruitfly) were sequenced soon after. The genomes of **human** [102,103], **mouse** [104], **dog** [105], **cat** [106] and **platypus** [107] all have been milestones for evolutionary and general biology.

To be sure, while sequencing of a representative organism provides the genome for that species as solid data, there is also much value in sequencing multiple individual members of the species of interest for studies of genetic variation within populations. This is especially so for *Homo sapiens*, and we will encounter this again in the areas of mapping of disease genes and pharmacogenomics below. And this in itself brings up the issue of how genomics itself relates to drug discovery. A quick answer is at the level of targets and defining them on a genome-wide scale, especially relevant from the point of view of the hazards of excessive target reductionism noted in *Searching for Molecular Solutions*. Placing a target into its genomic context is thus a sound pharmacological policy, and for this, detailed genomic information is the *sine qua non*. Just as a single target exists within a complex interactive network within the cell which encodes it, target molecules can also be thought of as members within a radiating evolutionary network. Mining of genomic data combined with evolutionary crosscomparisons between species can therefore be a powerful approach to pin-pointing likely entry points for drug targeting. Moreover, these kinds of analyses can directly lead to drugs themselves, in the form of natural products. The world of natural small molecules maps back onto specific genomes in the sense that all these low-molecular weight entities arise from enzymes, and knowledge of the genomically-encoded pathways involved can have valuable practical ramifications. A special off-shoot of genomics, metabolomics (also termed metabonomics) is concerned with this and related areas. Beyond information gathering, genomic engineering itself has become a powerful route for the production of certain useful biomolecules, and this trend is bound to increase.

We will examine some of these general aspects of genomics and molecular discovery in a little more detail soon, but there is another aspect of practical genomics which slots well within this section. If we look again at Fig. 8.Nf, there is an implicit message about technological development over time. Between the small genomes analyzed in the late 1970s (whose sequencing at that point was a considerable achievement) and the mammalian genomes sequenced in the early years of the 21st century, there is a size differential of almost six orders of magnitude [♥]. Clearly, this kind of progress could not remotely have taken place if the same technology as used initially was maintained throughout. The rise of genomics is then intimately associated with advances in sequencing technology, which is necessarily carried out in a high-throughput volume to cope with the size of large eukaryotic genomes. At a conceptual level, sequencing and chemical library screening thus intersect through their mutual need for high-throughput technologies, even though the specifics in each case are of course quite distinct and varied. While the details of modern high-throughput DNA sequencing need not preoccupy us here, we could note in passing that if a few key words or phrases had to be assigned to distinguish 21st century genomic sequencing from efforts in the 1970s, above all we would find ‘automation’ and ‘parallel

[♥] The smallest genome of Fig. 8.Nf is the plasmid pBR322 at 4.36 kilobases (4.36×10^3 bases), in contrast to the haploid human genome at approximately 3,300,000 kilobases (3.3×10^9 bases).

processing'. Also, while 'Sanger' sequencing ♥ was the method of choice for decades, in recent times new chemistries for DNA sequencing have been introduced. Both non-Sanger chemistry and parallel processing are very much evident in relatively recent techniques such as 'pyrosequencing' [18,110] and other technologies [19,20], which give only short read-outs for individual determinations, but triumph through massively parallel analyses ♣.

When one is talking of nucleobase sequences in the billions, the sheer volume of data would neutralize its usefulness if there was no efficient way to collate and analyze it. Here, of course, computational processing and the science of bioinformatics come into play, and these too have developed massively over the time period covered in Fig. 8.Nf. In published work studying genomes (or their products, as we will see more of later), the terms 'global' or 'genome-wide' are very frequently used as descriptors, but any analyses on a large genomic scale would be impossible without high-level computational assistance, as well as access to the primary sequence data itself.

Genomic analyses have thus moved from feasibility to demonstrably possible to almost routine, through a convergence of enabling technologies [111]. What was once breath-taking in its audacity, becomes routine as technology marches on. Vices (cost over-runs, unrealistic aims; grant funding denials) can thus turn into virtues (economies of scale, high-throughput genetic analyses, grant success). And the pace of these changes is relentless. As one more example, consider that the sequence of *Escherichia coli* (strain K-12) was published in 1997 [97]. This bacterium (and the K-12 strain of it in particular), is familiar to biologists world-

♥ This method, named for its inventor [108] is based on chain-termination with dideoxynucleotides, and generally has been more widely used than its contemporaneous rival technique of sequencing by controlled chemical degradation, the Maxam-Gilbert approach [109]. Both Sanger and Gilbert were awarded Nobel prizes in 1980 for their work in DNA sequencing.

♣ A little more detail on recent sequencing developments is given in the file SMS–CitedNotes-Ch4/Section 6 (PCR); from the same ftp site.

wide as a laboratory work-horse applied towards unraveling many features of biochemistry, genetics, and molecular biology for well over half a century [112]. In contrast, in the same year the genomic sequence of a pathogenic strain of *Helicobacter pylori*, the causative agent of peptic ulcers, was reported [113]. In this case the gap between the very discovery of the organism [♥] [115] to the sequencing of its genome was only 15 years, and this period coincided with the rising of the genomic age.

Targeting the Big Picture with a Genomic Eye - Searching for Targets with Homologous Sequences or Motifs

How can masses of genomic sequence data translate into new drug targets? A reasonably obvious way is to inspect the human genome for additional members of protein families which have already been useful in the target sphere. Pre-existing knowledge that a protein family has a proven track record in furnishing successfully druggable members, and that many more members are ‘out there’ as yet untapped, will act as an additional incentive for such computational searches. The success of such an undertaking is naturally based on the premise that previously-identified sequence motifs within known members of a protein family will be shared by others as yet unidentified, even across wide evolutionary gulfs. And there is plenty of precedent to support this supposition. Many evolutionarily linked groups of proteins have been usefully mined for drug targets, of which G Protein-coupled Receptors (GPCRs) are the most notable example. The latter in reality constitute a superfamily, such is their diversity of receptor function, although all share the core feature of possession of seven transmembrane segments [116]. The GPCR superfamily within the human genome has been accordingly subdivided into a limited number of families based on

[♥] This gastric organism had in fact been observed much earlier [114], but had not been cultivated or characterized. Its association with ulcers was based on an independent re-discovery event [115]. At 1668 kb, the genome of *H. pylori* is significantly smaller than *E. coli*, but still a formidable challenge by the standards of early sequencing.

sequence clustering, with each family named for a prototypical member with known function [116]. It was noted in Chapter 3 of *Searching for Molecular Solutions* that a large number of GPCRs are devoted to olfactory sensing.

Searching a totally defined genome for sequence homologies should in theory provide a comprehensive picture of all superfamily representatives which will be caught by this net-casting, and thereby yield previously unknown members which might serve as potential pharmacological targets. But this great power of genomics at the sequence level reveals at the same time a weakness. Or at least a weakness manifested through our current limitations of analysis of highly complex systems, such as mammalian genomes. In other words, genomic sequence data can rapidly reveal new gene family members, but by itself it will not provide high-level functional information, or even the protein's direct function in many cases. We will return to this theme shortly, but the upshot of this point is that genomic analyses have revealed numerous genes with unimpeachable sequence links to known superfamilies or specific families, but without known functional roles. Although these are genomic 'orphans', they have the prospect of finding homes as our knowledge base increases.

Of necessity, numerous GPCRs have been given places within the genomic 'orphanage', and this in itself illustrates that orphan status in itself is not an absolute condition. In fact, the truest of orphans are sequence-defined open reading frames which correspond to no known proteins, or 'ORFans' [117,118]. If one is conducting a genomic search based on certain well-defined sequence motif criteria, then by definition such unknown 'hypothetical proteins' will not turn up in the end results. Again with the example of GPCRs, sequence screens for entitlement to membership in the GPCR superfamily will by definition pull out candidates with certain specific features. The '7TM' (seven trans-membrane receptor) characteristic defined as universal for GPCRs [116] should thus be present in the entire genomic GPCR repertoire. Orphan GPCRs are then not at all ORFans, since they are already assigned as receptors with a specific structural feature, mediating signal transduction events. Their orphan status is then at the level of their function, and in particular the nature of the trigger for the signaling processes which they are presumed to mediate. A problem for deducing the nature of such stimuli for GPCRs is their sheer diversity, ranging from photons (light elicits signal transmission from the optical receptor rhodopsin), calcium ions, various small molecules, peptides and proteins [119]. Computational classification of orphan GPCRs has a foothold when homologies with a known ligand-binding GPCR class exist, but have often been stymied when such similarities are absent [120]. This is simply re-stating the observation that the more an orphan protein is out on a limb without perceived relationships to known groups, the more difficult it becomes to bring it back into the fold (so to speak).

Orphan receptors are not by any means restricted to our current example of GPCRs. Nuclear receptors are also major drug targets, and a large fraction of the nuclear receptor superfamily qualify for orphan status through ignorance of their natural ligands [121,122]. An important caveat here, though, is the realization that both GPCRs and nuclear receptors may include orphan members whose mode of action does not stem from conventional ligand binding in itself. Experimental evidence has suggested that some orphan GPCRs function in a regulatory manner through ligand-independent heterodimerization with other (ligand-binding) GPCR proteins [123]. Also, certain orphan GPCRs show ligand-independent (constitutive) signaling function associated with pathogenic mutations, viral expression, or viral induction of endogeneous genes [123]. Some nuclear receptors assigned as orphans appear to function through interactions with cofactor proteins rather than conventional small molecule ligands [124]. Collectively, these kinds of observations would suggest a modicum of caution towards making assumptions about the specific functions of proteins based only on their demonstrated membership within a large superfamily.

Display technologies constitute a pragmatic approach towards finding a binding peptide for any receptor whose natural ligand is unknown. For example, phage display has been used to define peptides which bind to the surface immunoglobulins of B lymphoma cells [125], towards which no ligand-binding information is otherwise available. Peptide 'mimotopes' mimicking a natural ligand can then in principle be found for many other receptors as well [126,127], although the success of a peptide mimic for a non-peptidic small molecule ligand may be limited. Yet a peptide ligand mimic showing some specificity for orphan receptor binding may be a useful handle for functional studies of an otherwise elusive signaling system.

Enzymes feature prominently as drug targets, and a surprisingly large set of them are orphans in a reverse sense to the above orphan receptor cases. Rather than genomic sequence orphans, many enzymes still exist only as described activities from biological samples rather than at the level of fully-defined proteins and genes [128]. Here we should recall (from Chapter 2 of *Searching for Molecular Solutions*) that the enzyme EC numerical classification scheme is based on the type of enzymatic process involved. An EC assignment then refers to a catalytic activity, independently of the biocatalyst mediating the observed effects (which in principle might even be a ribozyme rather than a protein enzyme, albeit an unlikely scenario in almost all cases). A significant fraction of all assigned EC numbers stand without an accompanying gene sequence in databases or the wider literature [129], and these may be a large untapped drug target source [128]. Although they are 'reverse-orphans' from a genomic point of view, modern genomics itself can be brought to bear in a concerted effort to fill this gap between described activities and molecular sequence information [130].

This 'known activity, but unknown gene' scenario for certain enzymes raises the more general issue of moving from functional biological observations to the gene level. Genetic diseases have historically presented phenotypes which were often very difficult to pin down at the molecular level prior to advances in molecular biology, since the physical nature of the defective gene products in most cases was completely unknown. Let's then take a quick look at the role of genomics in solving these problems and bringing the responsible proteins into the light of day as potential drug targets.

Finding targets in the first place through genomics

Just as the present age has seen an outpouring of genome sequences which can never be repeated in the same manner, the past three decades have seen a cascade of discoveries in defining disease genes which will also prove to be historically unique. Cloning of some disease genes was relatively easy owing to pre-existing biochemical information. For example, the genetic disease phenylketonuria was long known to be caused by a deficiency in the enzyme phenylalanine hydroxylase, which converts the amino acid phenylalanine to tyrosine. This allowed the cloning of the corresponding cDNA relatively early in the development of molecular genetics, and in fact before the assignment of the phenylalanine hydroxylase gene to a specific chromosomal site [131,132]. Yet for the majority of genetic diseases, no such information was available, and mapping techniques which made no assumptions regarding the nature of the culprit gene products had to be developed. Inaugurated with the discovery of the cystic fibrosis gene in 1989 [133-135], these advances have often been items for the popular press, which usually hail them as harbingers of treatments or cures. Despite the 'hype' which tends to shower over such reports, the essence of this is true and logically compelling. It is, after all, a matter of acquiring specific knowledge of a potential drug target, as a means for greatly increasing the options available for drug development, and greatly improving the odds for success. In the absence of any other information, a human genetic disease such cystic fibrosis is a formidable problem for drug intervention, since a natural animal model is absent. Definition of the relevant disease gene and its product allows not only concerted efforts towards correcting the defect by empirical screening or rational design *in vitro*, but also the deliberate engineering of highly specific animal models by gene targeting approaches. So while defining a disease gene and its product does not confer immediate therapeutic benefits, it is a giant stride in the right direction.

As with the conundrum of orphan genes, full genome sequence information does not automatically identify disease genes, but the problem is at a different level. By definition, a 'disease gene' is defective in some manner in comparison to a standard normal copy in the majority of the population. As such, it can only be defined by comparing genomic sequences between normal individuals and those stricken with the genetic disease under study. For a relatively simple arrangement where a disease is caused by a single defective gene, the correct disease gene should clearly show defects only in victims of the disorder [♥], and not in normals. In order to reach this point, complete genome sequence information is not necessary as long as some kind of *genetic markers* are available through which the gene can be mapped [♣]. Linkage mapping of genes through their observed phenotypes has long been performed through classical genetics, and is based on the frequency of recombination occurring through meiotic genetic interchange in eukaryotes. (As we have seen [*Searching for Molecular Solutions* Chapter 2], sex is very fundamental in biology, and *E. coli* genes were mapped long before rapid DNA sequencing, through bacterial sexual

[♥] A considerable amount of information can be gained from prior genetic analysis of the disease, to distinguish conditions which are manifested by a single genetic defect (and behave genetically as Mendelian traits) from those which are influenced by multiple genes (polygenic). Also, such analyses will reveal whether the disease is dominant or recessive (requiring two defective copies in the latter case) or sex-linked (present on the X-chromosome).

[♣] Note that in this section, 'genomics' is used in the broad sense of analysis of genomes, not necessarily involving a genome-wide survey or knowledge of an entire genomic sequence. Certainly, at the outset linkage mapping with chromosomal DNA markers can be carried out by genetically tracking the markers without any genomic sequence information. This has even been called 'hypothesis-free' biology, but of course the hypothesis is actually that the origin of an inherited Mendelian disorder lies within the genome itself, and that linkage mapping can genetically localize it [136]. This hypothesis has been amply validated.

conjugation). Yet for human genetic analyses, this approach is very limited, and practical mapping of disease genes requires physical chromosomal markers which are genetically polymorphic (represented by alternative forms within the population).

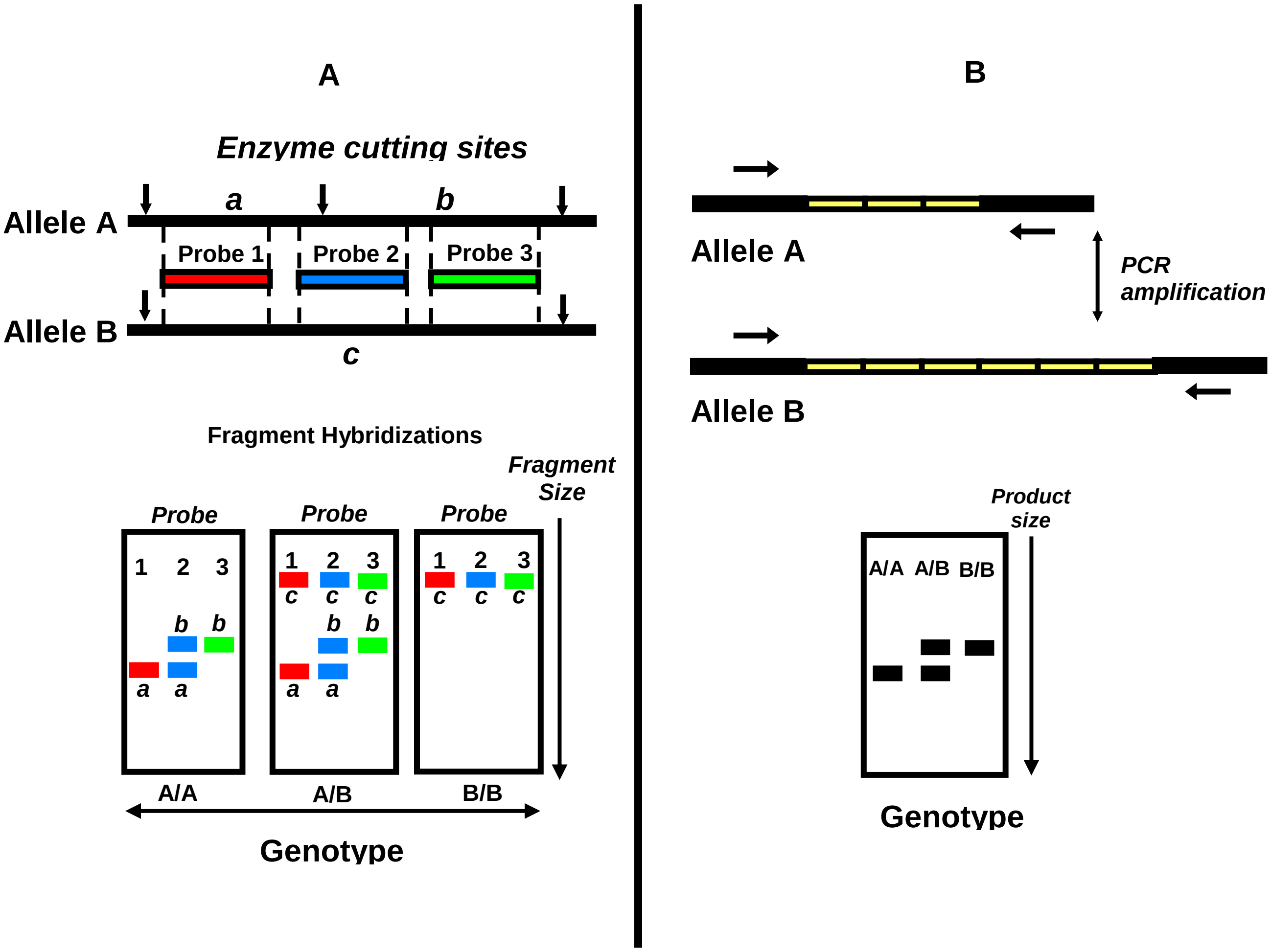


Fig. 8.Ng

Methods for analysis of genomic polymorphisms. **A**, The early restriction fragment length polymorphism (RFLP) approach. Two alleles of a genomic locus are depicted, where an RFLP exists for an enzyme cutting DNA at a specific site (vertical arrows; polymorphism corresponding to middle arrow). The polymorphism is analyzed by hybridization with specific complementary probe sequences (1-3 as shown; red, blue and green bars respectively). Fragments are separated by size (electrophoresis) and hybridizing bands are detected (Southern blotting). The banding pattern is dependent on the sequence spanned by the chosen probe and the genotype of the sample. (A/A and B/B are homozygotes of alleles A and B; A/B is a heterozygote). Probes 1 and 3 are localized within restriction fragments and only 'light up' bands *a* and *b* respectively, while probe 2 spans the central site and hybridizes to both fragments. **B**, Polymorphism for a short tandem repeat analyzed by polymerase chain reaction, with alleles distinguished by size differences on an electrophoretic gel.

At the beginning of the 1980s, it was recognized that polymorphisms of arbitrary chromosomal DNA fragment lengths produced by restriction enzymes (restriction fragment length polymorphisms, or RFLPs) could serve as very useful chromosomal markers [137] (Fig. 8.Ng-A). RFLPs which segregate with a disease phenotype within an extended family pedigree reveal the chromosomal location of a disease gene, as was shown early on with screens for polymorphic markers for Huntington's chorea [138]. If a polymorphic marker showing tight linkage with a disease trait can be identified, a relatively small chromosomal segment bearing the gene of interest can be focused on. (Although here 'relative' refers to the size of entire chromosomes, and the narrowed-down area might still encompass millions of base-pairs). At that point, physical cloning methods and chromosomal 'walking' can be applied to systematically search for the desired gene, as was first done for the above-mentioned cystic fibrosis gene. After the advent of the polymerase chain reaction, RFLPs were largely superseded by amplification of polymorphic short tandem repeat sequences (Fig. 8.Ng-B; also known as 'microsatellites'). These elements are widely (and more or less randomly) distributed throughout the genome and are convenient markers for genetic mapping purposes, including on a genome-wide basis [139]. But the most common type of genetic variation is also at the most basic genetic level, that of *single nucleotide polymorphisms* [♥] (SNPs, and pronounced as ‘snips’) [140].

The success of pin-pointing the specific mutational origins of many genetic diseases in the 1990s generally hinged upon the relative simplicity of the specific diseases which were targeted. Once the basic techniques for physical chromosomal linkage analyses had been established, in principle it was possible to define any genetic disease whose origin was traceable to a single gene, given access to DNA samples from members of extended families harboring the defective gene in question. Most of the single-gene ‘Mendelian diseases’ have in fact yielded their secrets to this form of systematic investigation. In the afterglow of these heady successes, many workers began to think of ways such genetic knowledge could be applied with a therapeutic bent. The clinically important defective genes bear mutations which ablate or decrease function of the corresponding encoded proteins, in the form of nonsense or missense codons, or generation of aberrant splicing events. To overcome such loss-of-function mutational impairment, a conceptually simple but practically difficult solution is to provide and express a normal gene copy, the goal of *gene therapy*. This large area, which has progressed significantly from its inception, falls somewhat outside our ambit, although it certainly constitutes a ‘molecular solution’ of sorts when successful. But in any case, there are other feasible solutions. Ribozymes have been shown to be capable of remarkable feats of mRNA repair, of potential great significance for many clinical purposes, noted in Chapter 9 of *Searching for Molecular Solutions* [♣]. It might be at first thought that small molecules have no place in therapies for genetic diseases, but this view appears too pessimistic at

[♥] In practical terms, a single-nucleotide polymorphism is ‘common’ and analyzed within the HapMap project (noted shortly in this section) if both alleles occur at a frequency of ≥1%, and common SNPs account for 90% of single-nucleotide variation. A very large number of rare alleles account for the remaining 10% of human variation [140].

[♣] Another relevant technology referred to in Cited Notes for Chapter 4 (SMS-CitedNotes-Ch4/Section 8B; from the same ftp site) is genome engineering by zinc finger nucleases.

least for a specific subset of cases. Mutations which result in nonsense codons (translational termination signals) cause premature halting of protein synthesis and production of defective truncated polypeptides. Certain aminoglycoside antibiotics have the effect of causing the translational machinery to over-ride such unwanted signals, and continue to synthesize full-length protein. Pilot studies with aminoglycosides such as gentamicin in appropriate patients (with nonsense mutations in their defective genes under treatment) produced mixed results [141,142], but these data inspired further searches for improved agents, especially since reinstatement of as little as 5% of normal protein levels may greatly ameliorate symptoms in at least some genetic diseases [143]. A subsequent commercially-derived drug (chemically unrelated to gentamicin; PTC-124) has shown greater efficacy as a termination suppressor and considerable promise in animal models [143,144]. Of course, this kind of therapy is limited to the subset of patients with this specific class of mutation, which can vary widely depending on the disease gene in question [143]. It is nevertheless a very significant area of continuing investigation and clinical development, and an exemplar of how information from the genome can flow back into the world of useful small molecules.

Though of great biological and clinical significance, the identification of Mendelian disease genes for the most part concerned rare conditions in the general population, and did not address the genetic influences for more common diseases. Rather than arising from single genes with clear mutational lesions at the DNA sequence level, many diseases arise from a complex interplay between natural gene variants and environmental factors [♥]. New drugs capable of positively modifying such complex and common problems as Type II diabetes, obesity, mental illnesses and various autoimmune diseases might acquire rapid blockbuster status. Once again, definition of the genetic background to such

[♥] An example noted in the file SMS–Extras-Ch3/Section A3 (Autoimmunity) is the autoimmune disease multiple sclerosis.

conditions is of obvious importance in development of ultimate solutions, however complex the genetic interplay may prove to be.

A genome-wide database of human SNPs has considerable potential for assisting studies attempting to draw genetic associations with complex diseases. To this end, since 2002 an international consortium has amassed human SNP data for the 'HapMap' project (Hap certainly not as in hapless, but *haplotype*, which corresponds to the specific pattern of SNP alleles for a single chromosome, or part thereof [140]. Genomic haplotype maps of increasing refinement have been duly delivered, incorporating millions of defined SNPs [145,146]. Haplotypes can change through new mutations arising, or meiotic recombination, but are not assorted randomly by the latter. Throughout the genome, it has become evident that haplotypes are patterned as distinguishable blocks, sizable tracts of chromosomes within which recombination is low [147-149]. This is the outcome of a genetic effect termed *linkage disequilibrium*, which is present when two genetic alleles do not segregate in an independent manner [♥]. Linkage disequilbrium and resulting haplotype blocks have value for large-scale genetic association studies [148,149,152], and also for studies tracing the origins of modern humans [153-155].

SNP data generated through the HapMap project are amenable for application in genome-wide association studies in a high-throughput manner, through commercially produced 'gene chips' (microarrays) detecting large panels of

[♥] In a population, if two alleles (A and B) at two genetic sites (loci) are independent and found at frequencies *A* and *B* respectively, the frequency of the combined AB haplotype is the product *AB*. But if the observed frequency is either significantly higher of lower than this, linkage disequilibrium exists [150]. In regions of very high linkage disequilibrium, alleles thus tend to occur together and are separated by recombination at rates lower than expected if purely random processes were operating. Such genomic regions of high linkage disequilibrium, bearing specific constellations of alleles of low diversity, are accordingly termed haplotype blocks [147,148]. Boundaries of haplotype blocks correspond to recombinational hotspots [146,147], which collectively account for a high proportion of meiotic cross-over events [146,151].

SNPs via hybridization techniques [156]. Many confirmed associations between specific genes and complex diseases have been made in this manner, including obesity, diabetes, and a variety of degenerative conditions [156-158]. (The list is long and explosively growing; a catalog of such studies is maintained by the National Human Genome Research Institute / NIH). It might be expected that many (or most) SNPs associated with disease would represent nonsynonymous mutations within coding sequences, resulting in variant (and possibly suboptimal) proteins. Yet the SNP-harvesting experience has not borne this out, with most changes associated with increased disease risk occurring outside coding sequences [156]. This in turn highlights the profound importance of gene regulation in the complex interactive networks operating in a multicellular organism.

But in many cases, the association between a gene and a pathological condition is indicative of a small increased risk (<1.4 fold [159]), and where multiple genes are so implicated, untangling the complex routes of the pathology will take concerted time and effort. Polygenic contributions to any phenotype (not merely complex diseases) are another type of combinatorics in action (a subtheme of *Searching for Molecular Solutions* Chapter 10), and the complexities arising from the interplay between multiple genes has been termed the 'problem of dimensionality' [160]. This refers to the rapidly escalating number of combinatorial possibilities arising as the number of potential interactive gene combinations increases, even if the human genome 'only' has on the order of 20,000 genes. Phenotypic effects of a genetic allele in multicellular organisms can indeed be strikingly dependent on genetic background [161]. One way of reducing such complexity is to direct the genetic screening focus towards genes whose products are known to mutually interact [160] (such information itself may be obtained through genome-wide functional screens, noted in Chapter 9 (and its associated Cited Notes; from the same ftp site).

But not all polygenic diseases have readily yielded convincing information from genome-wide SNP association screens [136], raising the question as to whether a considerable portion of specific disease burden is not necessarily linked with genetic variants which are common in human populations. This has been framed as the ‘common disease / common variant’ (CD/CV) hypothesis, as opposed to the scenario where diverse rare variants influence common diseases. It has pointed out that these proposals are not necessarily mutually exclusive, depending on specific disease states [162]. Apart from issues relating to the accuracy of scored associations and sheer complexity, there are additional interesting reasons why genome-wide association screens may yield incomplete data (or fail to register) for a subset of complex diseases. Screening only for changes at the DNA sequence level (as with SNPs) will not take into account epigenetic modifications [♥] (Chapter 2), and these are increasingly becoming associated with complex disease phenotypes [163-165]. Since even genetically identical organisms (as with identical twins) can show phenotypic variation association with epigenetic differences [166,167], it becomes less surprising to find roles for the ‘epigenome’ in pathological states as well as normal control processes such as genetic imprinting (Chapter 3). For some disease states (such as some forms of epilepsy), somatic rather than germline mutations have been suggested as having pathological significance [168]. Still, even if a valid observed genetic association with disease is rare (and thus not useful for general screening), genome-wide association data is invaluable for the identification of new drug pathways and specific targets of generalizable utility [156,169]

There is one disease state where both epigenetic and somatic mutational changes are not at all controversial, and that is cancer. High-throughput genome-wide studies of cancer epigenomes have been undertaken, with many investigations continuing to reveal epigenetic alterations in the pathological transformed cell populations [170-172]. The search for somatic mutations in cancer cells has been launched on a grand scale. One such initiative based in the UK, the Cancer Genome Project, aims to screen ultimately all human genes for

[♥] Genome-wide array-based screen can be adapted to screen for epigenetic cytosine methylation also [163].

somatic mutations driving cancer cell growth and survival. Initially concentrating on sets of genes regarded as more likely to be significant in driving cell proliferation, the project rapidly uncovered a very significant mutation in the signaling kinase B-raf [173], an important somatic event in a number of tumors, including melanoma. A US program, the Cancer Genome Atlas, was initiated in late 2005 with ambitious sequencing aims [174], subsequently modified towards an early focus on mutations in known cancer genes [175]. It is generally agreed that ultimate success of these large-scale projects is dependent on processing of a large number of tumor samples, to reliably distinguish the noise of ‘passenger’ somatic mutations from ‘drivers’ which actively assist aberrant tumor cell growth [176]. Obtaining access to the desired number of samples is in itself a formidable logistical problem which may limit or retard ongoing progress of these ‘big science’ undertakings [177].

At Home with ‘Omics

In addition to the major ‘omics considered in Chapter 9 of *Searching for Molecular Solutions*, the ‘high-dimensional biology’ [178] of the genomics area has given birth to a plethora of additional ‘omics in a seemingly ever-branching cascade. These can be named and split according to one’s field of interest, as long as they are pursued on the necessary global scale. As one example among many, genomics as applied to the area of biological toxins has been termed ‘toxicogenomics’ [179,180], and a subset of this in turn which has given the catchy label of ‘venomics’ [181] (a portmanteau word for the genomic era). Just as the Watergate affair bequeathed the –*gate* suffix to the language for affairs only tangentially related to the original, the meaning of ‘omics may further diversify [♥], although purists will doubtless object. Perhaps inevitably, the many new ‘omics have been collectively dubbed ‘polyomics’ [183], not to be confused with its

[♥] As one example of this, the application of fragment-based chemistries (as considered in Chapter 8 of *Searching for Molecular Solutions*) has been termed ‘fragonomics’ [182], although its relationship to genomics is very peripheral at best.

anagram 'polysomic' (bearing additional chromosome copies above the norm) or polymer analyses.

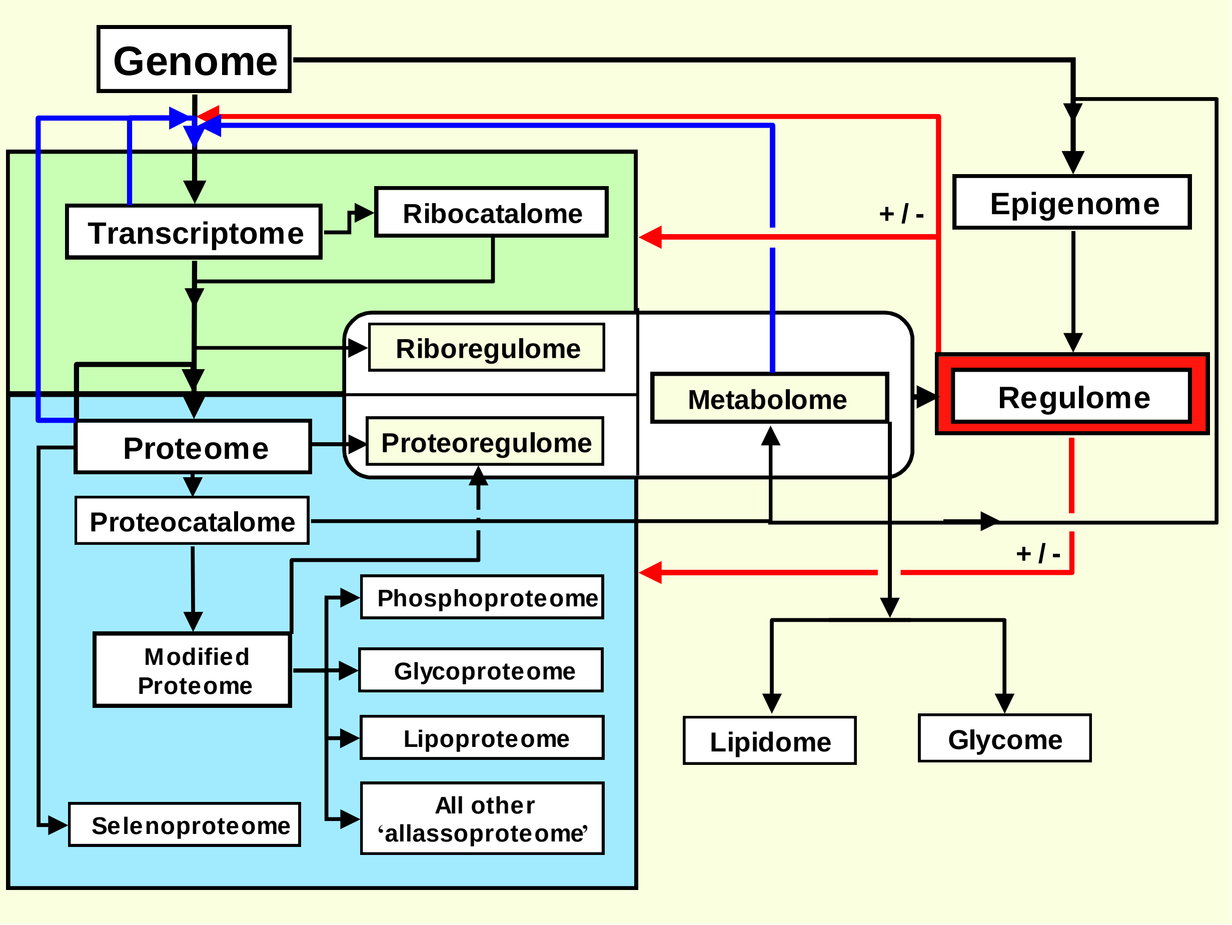


Fig. 8.Nh

Genomic networks in mammalian cells. Directional lines: **black** lines and arrows indicate 'outward information flow' encoded by the central genome, which gives rise to the machinery (contained within the transcriptome and proteome) and the materials (contained within the metabolome) for genomic replication itself, shown by **blue** lines and arrows flowing back to the genome. The transcriptome, proteome, metabolome and epigenome contribute to the regulome, which positively or negatively regulates genomic replication or its expression (shown by **red** lines flowing to the genome, proteome, and transcriptomes). The **regulome** is defined as all genome products (direct or indirect) involved with genomic regulation (in a broad sense which is taken as including the epigenome itself). Subsets of the regulome are the riboregulome and proteoregulome, involving RNA-based (such as RNAi and miRNA) and protein-based (such as transcription factors) regulatory mechanisms, respectively, although these overlap in certain areas. Other designations as follows: ribocatalome, ribozymes encoded within genome; proteocatalome, all protein enzymes encoded by genome; modified proteome ('allassoproteome', from the Greek *allasso*- to alter), subset of proteome bearing post-translation modifications; lipidome and glycome, small lipids and carbohydrates synthesized by protein enzymes. The selenoproteome (all proteins containing selenocysteine) forms a special category since it makes use of selenocysteine as a 21st amino acid (as noted in Chapter 5 of *Searching for Molecular Solutions*). The ribocatalome is known to enable the proteome through ribozyme activity in ribosomes. Not that for simplicity, not all possible inter-relationships are shown. For example, the proteoregulome subset of the proteome can feed back upon the transcriptome by degrading RNA transcripts. (Hence a general red-arrow pathway from the regulome to both the transcriptome and proteome is used). Compare this chart with the cyclic depiction as in Fig. 9.7 (and the corresponding color version from the same ftp site) of *Searching for Molecular Solutions*.

This is all a matter of lumping or splitting of categories, as shown in Fig. 8.Nh. For example, the subset of the proteome performing catalysis (enzymes; proteocatalome) can be further split into the kinome [♥], or the subset of enzymes which function as kinases (transferring phosphate groups, of major interest as drug targets [184]. The complement to the proteocatalome is the ribocatalome (ribozymes), which is vestigial compared to its heyday during the RNA world, but still vital for modern organisms within the ribosome – and maybe in other important ways not yet fully defined. Non-covalent interactions within the

[♥] Note that in the terminology of Fig. 8.Nh, the kinome is a subset of the proteocatalome, and the phosphoproteome (the global cellular content of phosphate-modified proteins) is the catalytic *product* of the kinome. There is considerable overlap between the kinome and the phosphoproteome, though, since much signal transduction flows by phosphorylation of one kinase by another, and many kinases are self-phosphorylating (autocatalytic).

interactome are fundamentally important (as with protein-protein binding events), but so too are covalent chemical modifications, mediated through the agency of the proteocatalome, sometimes in interplay with a subset of the transcriptome and metabolome. This results in the epigenome and all chemically-modified proteins (dubbed the 'allassoproteome' in Fig. 8.Nh).

More will be added to the topic of 'omics in Cited Notes for Chapter 9 (see the file SMS–CitedNotes-Ch9/Section 30). A general point to note is that grouping subsets of genomic information into a designated 'omic category can occur at a range of different conceptual levels. At the primary level of converting genomic information to RNA molecules we thus find the transcriptome, but it is possible to make 'omic groupings of proteins and / or RNAs which contribute towards specialized functions at far higher stages of complexity in multicellular organisms. A good example of the latter is the coining of the term *lexinome*, in reference to the subset of the human genome which contributes to the high-level neurological systems specifying and enabling language and reading ability [185].

But from the special point of view of small molecules, there are some special 'omics to consider. Let's get there via a brief metaphorical introduction as follows, using a character from *Searching for Molecular Solutions*….

A Genomic Diversion

> In her ongoing quest for molecular enlightenment, Lucy has gained an audience with a biotechnological savant in much more familiar surroundings than some of her previous escapades.
>
> "We'd welcome the chance to sequence an Australopithecine genome," said the scientist. "It would be a rare opportunity to fill in some important gaps in the evolutionary record for primates. Neanderthals were one thing, but getting halfway decent samples for PCR from multimillion year-old humanoid fossils is

another thing entirely. At least for anything like complete genomic coverage. You can forget about Jurassic Park scenarios, you know."

Some of the biotechnologist's allusions were quite obscure, but Lucy persisted, ready if necessary to trade a blood sample for knowledge. Information for information in the end, she thought. Although Lucy's ability to speak was poor, in her inscrutable way she communicated her desires to the scientist. Lucy wanted to know something about the history and future of genomics, especially with respect to molecular discovery.

The scientist responded, "At the outset there were many who thought of genomics as nothing more than run-of-the-mill DNA sequencing writ large and with a big budget, and more than a little over-glorified to boot. But this somewhat myopic view did not take into account that full sequence information is necessary for a complete systems view of the organization and control of the expression of the genome, which is a wonderful thing. While we haven't yet got the full picture of *your* genome, we can have total confidence that the differences between you and me come down to a relatively tiny fraction of our net genomic sequences – it's all about developmental controls, and a few key mutations influencing language and mental development.

"Anyway, 'genomics' itself was only the tip of the iceberg. The trend for 'high-altitude' views in biology at all levels really caught on, and was made practical by converging enabling technologies. Soon there were other 'omics' coming out thick and fast – proteomics of course, but also epigenomics, transcriptomics, metabonomics….I lose track. Though all were meaningful and scientifically valid, there was clearly a popular trend in this direction. I'm surprised the first draft of the chicken genome wasn't hailed as ushering in the new era of 'henomics', but it doesn't seem to have caught on…."

Noting Lucy's impassive expression, the scientist wondered about Australopithecine cognitive limits and capacity for humor, while Lucy wondered about the human capacity for self-deception. It was important to get things back on track.

“Ah, yes, molecular discovery,” said the scientist. “This comes under the heading of *chemogenomics*, which is really an interface between the chemistry of small molecules and the entire genome as a source of drug targets”.

The scientist then waxed eloquent in explaining how chemogenomics had ramped up the discovery rate for pharmacologically beneficial small molecules, and described some genome-wide ‘tricks’ which facilitated scanning for useful chemical modifiers.

“But anyway, soon we’ll have your genome, and that may even impact on drug discovery, in the area of pharmacogenomics…”

Lucyomics?

“Well, you could put it that way, I suppose – people facetiously referred to ‘Watsonomics’ and ‘Venteromics’ after some early individual genome sequencing – I don’t recall the details, it’s all lost in the mists of time now. I do remember calls for working up a genomic record for past US presidents – people do get carried away – you know, along the lines of Washingtonomics, Nixonomics, Reaganomics….as if that were possible….. But individual genome sequencing itself is a dime a dozen now, of course.”

Costs?

“Not a problem at all, they fell very sharply early on, leading to genomes as cheap as chips, so to speak. A matter of econ-omics, you might say…”

Lucy’s genomic quest was hardly complete, but time had run out, and the much-prized blood sample was called for.

‘Omics to OMICS

Living systems cannot function without a multitude of small molecules, which collectively slot into the metabolome (sometimes also known as ‘metabonome’). This special area of study for natural small molecules overlaps with two other ‘omics or great relevance to drug discovery and practical optimization. The relationship between these, chemogenomics and pharmacogenomics, and metabolomics is depicted in Fig. 8.Ni. All three are concerned with small molecules, but while the ambit of metabolomics is the global set of small molecules present within an organism, the other two are not necessarily constrained to natural molecules as such, but interface with genomics at the level of targets.

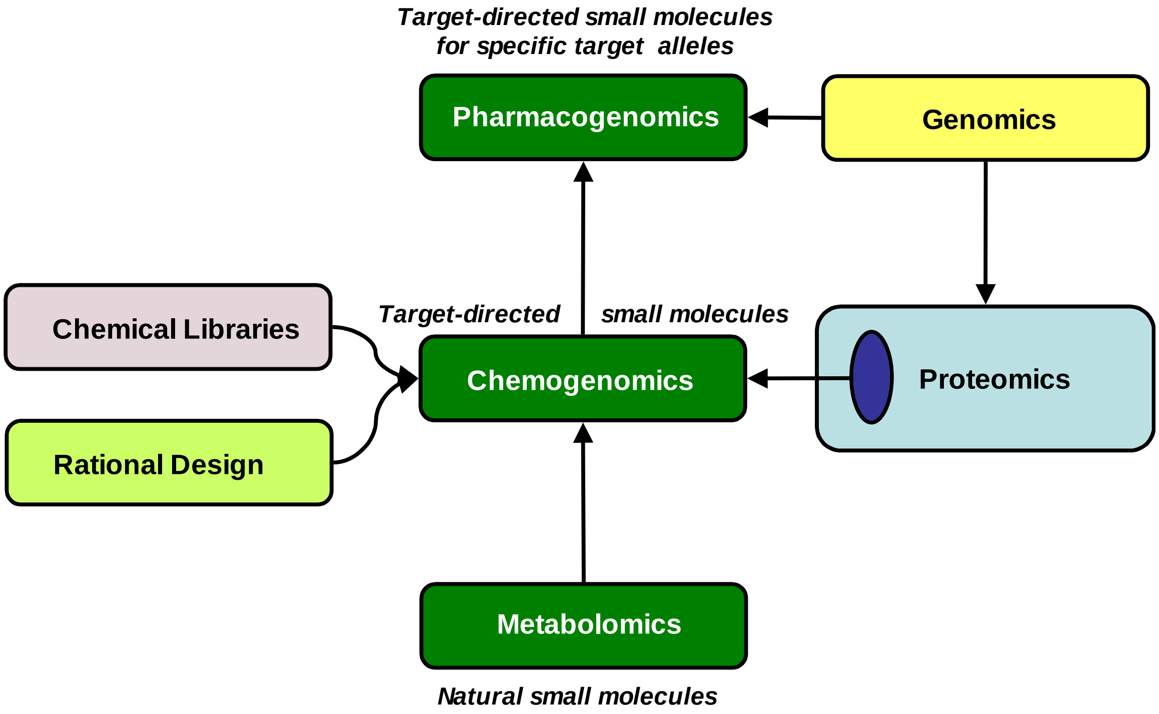


Fig. 8.Ni

Fig. 8. Ni. Relationships between metabolomics, chemogenomics and pharmacogenomics. Metabolomics is concerned with defining all small molecules present within an organism [186], while chemogenomics aims to systematically identify small molecules acting on the products of genomes [187]. For this, natural products (metabolomics) may be relevant, but input from chemical libraries and rational design is also important. Usually chemogenomics employs analyses of small molecules interacting with protein families (blue oval within Proteome), although RNA (transcriptome) and direct genomic (DNA-binding) targets are also possible. If chemogenomics contributes drug candidates, a further level of refinement for their practical deployment is pharmacogenomics, which aims to define individual variation in drug responses, and tailor patient-drug treatments accordingly.

Chemogenomics could be encapsulated as aiming at OMICS (Obtaining Molecules In Chemical Space [♥]) through 'omics, where the 'communication' between the these domains is a two-way street. This is so since the patterns of interactions of small chemical compounds with biological targets and systems provide information enabling the classification of the biological systems themselves, and invokes a type of complementarity between chemical space and a 'biological space' defined by phenotypic descriptors [188]. In fact, chemogenomics incorporates a number of themes noted within *Searching for Molecular Solutions*. One of is the need to move away from excessive target reductionism and consider multiple drug interactions ('polypharmacology') in the context of target families. In this context, an aim of chemogenomics is to map chemical ligand and target interactions into an integrated ligand-target space, which necessitates moving beyond a single-target focus [189]. The shift to a focus on target classes and families is fundamentally important in chemogenomics [189,190], and is effectively summarized by the principle that structurally similar target receptors will bind structurally-related compounds [191]. This was initially framed as the 'SAR (Structure-Activity Relationship) Homology' concept [192],

[♥] OMICS also tends to recall the more diffuse notion of OMspace, as noted in *Searching for Molecular Solutions*.

where targets show SAR homology if they bind structurally related ligands. Genomics enters this picture by providing global target data from which the related families can be identified, as we also have previously considered. Sequence motifs in known proteins conferring membership of a SAR homology group then can be searched in a global context [192]. (Druggable targets in general are usually proteins, but RNA or DNA targets may also be relevant (Fig. 8.Ni)). In fact, the emphasis of chemogenomics on drug discovery through the analysis of ligand interactions with target gene families is often seen as definitive for the term itself and separable from ‘chemical genomics ’[♥] [193], although this distinction is not always rigorously applied.

Chemogenomics can be practically implemented at the level of screening by cloning and expressing desired target families based on shared structural motifs [193]. Also, along the lines of the ‘quality over quantity’ principle for chemical libraries, increasing attention has been invested into ‘focused’ or target-directed libraries [189,194-196]. The notion of ‘privileged structures’ in binding ligands for specific target structures (Chapter 8 of *Searching for Molecular Solutions*) is inherently wrapped up into these objectives [197]. Chemogenomics is also becoming increasingly combined with investigations of model organisms with defined genomic alterations.

[♥] Chemical genomics or chemical genetics strictly refer to the analysis of biological systems by means of small molecules as agents to perturb normal function in an informative manner [193].

Section 23: ***Tethering***

Cited on p. 296 of *Searching for Molecular Solutions*

Molecular Discovery at the End of Your Tether

By definition, molecular targets must be present for the success of any screening process, even if the ultimate target is a poorly-defined component of a complex system. But targets have the special potential in fragment-based screens of serving as templates for the assembly of drug leads themselves. This is the case in the following section also, but here we will be concerned with strategies using physical linkages of fragments, or *tethering*, to direct fragment combinatorics.

In Chapter 8 of *Searching for Molecular Solutions* we noted how the principles of DNA-template directed synthesis reinforce the importance of local high concentrations in driving chemical reactions. This is relevant to non-covalent interactions as well, and can be applied in fragment-orientated combinatorial ligand screening. If a target (usually a protein) possesses a site where a readily reversible chemical bond can be instituted, this can be used as an anchoring point for molecular fragments to access non-covalent interaction sites within the rest of the target molecule. Disulfide bonds are a convenient means for such an attachment-mediated approach, and some proteins may offer native cysteine residues which can be exploited for such ends. This kind of process is termed site-directed ligand discovery, or Tethering®[♥] [198-200], as schematically depicted in Fig. 8.Nj below.

[♥] Tethering is a Registered Trademark of Sunesis Pharmaceuticals, Inc., and thus distinguishable from ‘tethering’ in the garden-variety sense.

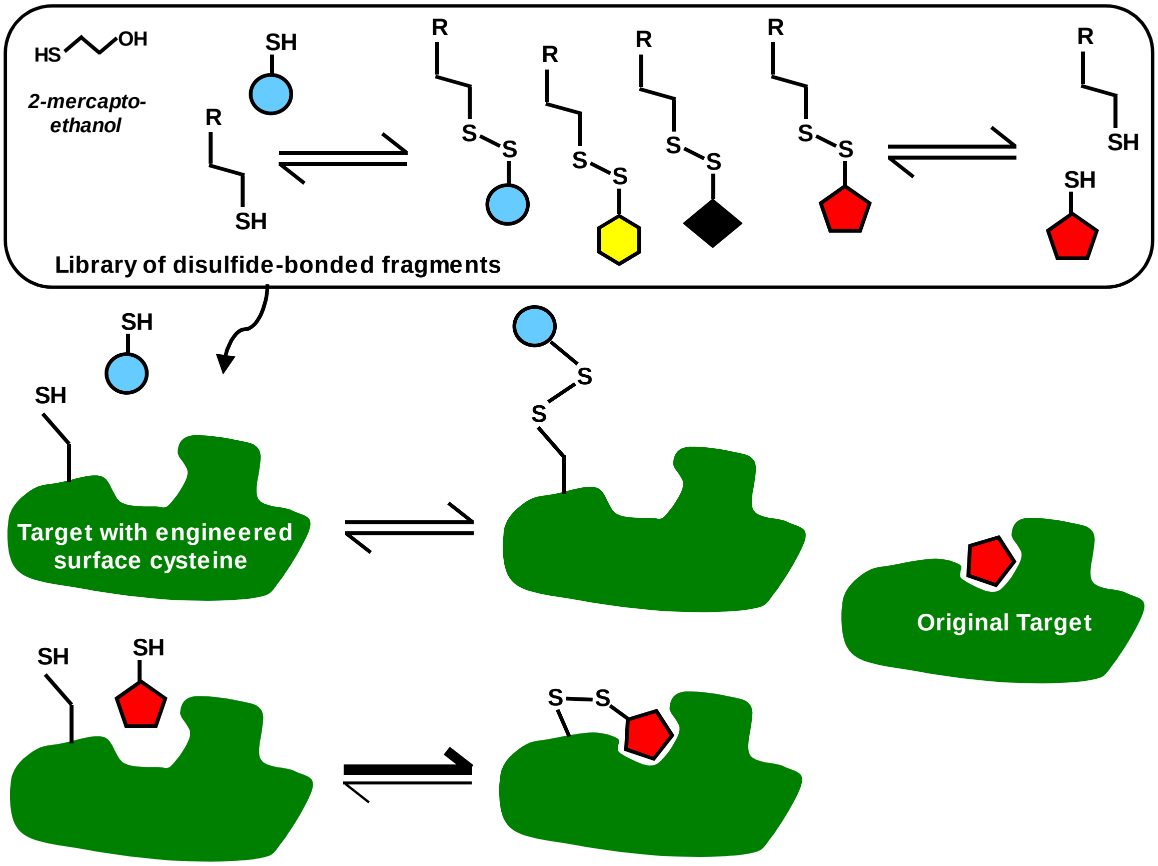


Fig. 8.Nj

Fragment tethering using a target with a cysteine residue introduced at a convenient site such that its sulfhydryl group is surface-exposed. A fragment library is constructed attached to a constant moiety, containing disulfides which interchange with their reduced forms (controllable by adjusting levels of reductant 2-mercaptoethanol). Disulfide interchange between the target surface sulfhydryl and reduced library elements reaches an equilibrium, but if the attached fragment interacts non-covalently with an accessible target site, the equilibrium is shifted to the right as shown, towards the target-bound product. Within the entire fragment population, the subset capable of target binding is thus ‘selected’ by the target itself and enriched on its surface. Following identification of fragments which bind target in this manner, the unmodified original target should still bind the free fragment as indicated.

Of course, surface accessible cysteine residues in proteins will rarely be so conveniently located, but cysteines can be placed at will within a primary sequence via standard site-directed mutagenesis. Other chemistries besides disufides are also possible [201], which could exploit the site-directed incorporation of unnatural amino acids (Chapter 5 of *Searching for Molecular Solutions*). The high local concentration of surface-attached fragments in effect drives the interaction with target sites by specific library members which can act as target ligands. This in turn drives the reversible sulfhydryl-disulfide equilibria towards persistence of the surface attached ligands (Fig. 8.Nj), which are thus ‘selected’ by the nature of the target itself [199]. The reversibility of this site-directed ligand discovery process is very useful, since irreversible anchoring of ligands has the danger that attachment may occur indiscriminately without assistance from non-covalent ligand binding [201].

A number of useful variations on the tethering theme have been developed [201]. One such process involves a kind of piggy-back strategy for fragment assembly (Fig. 8.Nk below). If an initial fragment binding a target site is irreversibly anchored *in situ*, and possesses an activatable sulfhydryl group, then the ‘selection’ process from a disulfide tethering library can be repeated to search for ligand binding to proximal sites [199,202]. Identification of such companion fragments allows rapid design of a stably-linked fragment pair, since the length of the disulfide arm is known at the outset [202]. This fragment-building approach can be applied towards obtaining novel pharmaceuticals [199].

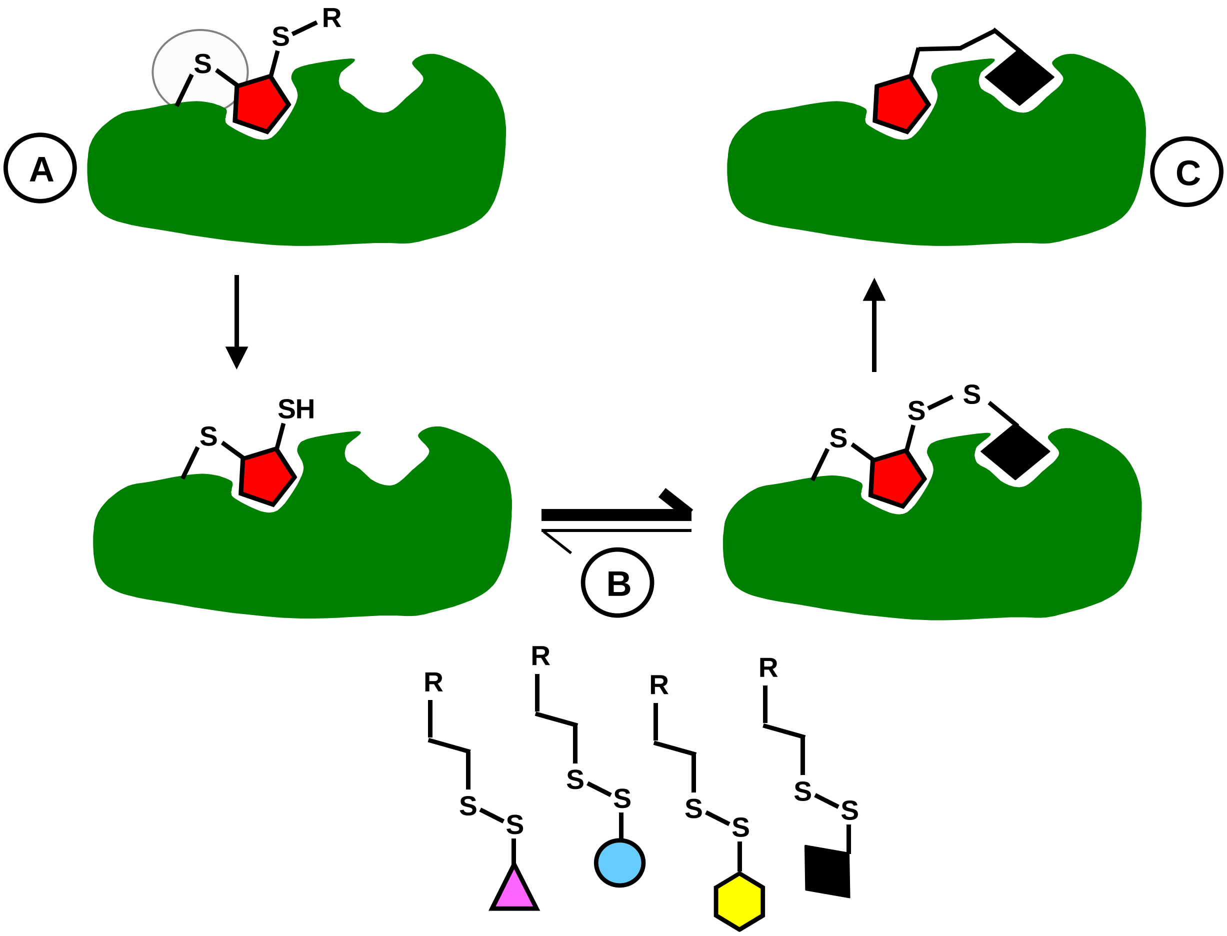


Fig. 8.Nk

Fragment assembly by tethering process. **A**, An introduced surface-accessible cysteine thiol is irreversibly modified (light gray circle) with a linked previously identified fragment bearing a masked sulfhydryl group. **B**, Chemical unmasking of the thiol group allows selection from a tethering disulfide library (as in Fig. 8.Nj; above) for a fragment binding to a proximal target site. **C**, The identified fragments are assembled with a stable linker of appropriate length.

In Chapter 8, it was also noted that phage ‘chemical display’ can be applied for the coding of libraries of small molecules. There is also a place for phage display in an analogous approach to the above tethering fragment assembly strategies [203]. Here the phage express on their surfaces a random peptide library fused to a protein domain which can form a specific heterodimer with another domain [♥]. If the latter domain is tagged with a specific ligand for a target of interest and provided in *trans*, heterodimerization allows selection for high-affinity binding, and can improve the specificity of resulting derived bivalent molecules (Fig. 8.NlA and B, below). In this strategy, the initial affinity of the ligand is a significant factor in the success of selection, as very tight binding by ligand alone will interfere with peptide selection. To circumvent this, the binding affinity of the primary ligand can be adjusted by using suitable analogs with reduced affinity [203].

[♥] The interaction between Fos and Jun were used for this purpose [203]. These proteins interact via a protein motif termed the leucine zipper, and form a coiled-coil structure [204,205].

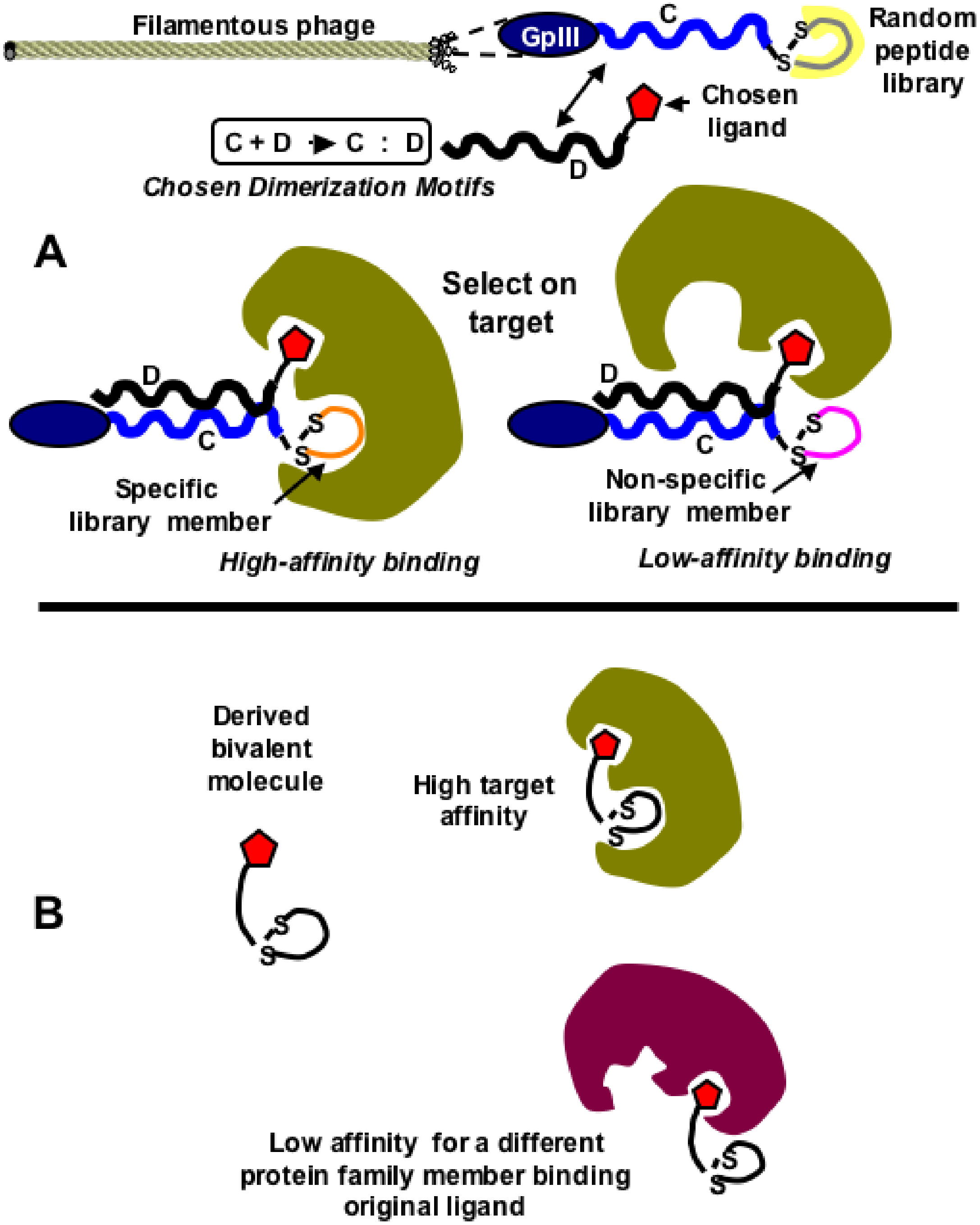

Filamentous phage
GpIII
C
S
S
Random
peptide
library
Chosen
ligand
C + D ► C : D
D
Chosen Dimerization Motifs
A
Select on
target
D
C
S
S
Specific
library member
High-affinity binding
D
C
S
S
Non-specific
library member
Low-affinity binding
B
Derived
bivalent
molecule
S
S
High target
affinity
S
S
Low affinity for a different
protein family member binding
original ligand
S
S

Fig. 8.NI

Fig. 8. NI. A, Phage display for co-presentation of a target-binding ligand to increase binding affinity and specificity [203]. A display library on the filamentous phage gene III product (gp3) consists of a protein domain C which specifically interacts with domain D provided in *trans*. The domain D also has a molecular fragment of choice appended to it. A peptide library (topologically constrained by cyclization via a disulfide linkage) is fused to the N-terminus of domain C. When the expressed phage library is exposed to the domain D / ligand, heterodimerization occurs, allowing selection for peptides with affinity for the target in tandem with ligand (the combined ligand / peptide binding results in selectably higher affinity than binding to ligand alone. **B**, Design of a linked bivalent molecule from display results, which has increased specificity for the original target over protein target family members which bind the same ligand but do not share the peptide-binding site.

Section 24: ***Encoded Self-Assembled Combinatorial Libraries***

Cited on p. 299 of *Searching for Molecular Solutions*

Another target-driven approach which also features self-assembly has been developed, termed ESAC for 'Encoded Self-Assembling Chemical Libraries' [206], as portrayed in Fig. 8.Nm below. Oligonucleotides appended to a small molecule library mediate self-assembly (through common sequences) and selected binding molecule identification through library element-specific tags. In a similar manner, a known low molecular weight target ligand can be used to search for additional co-binders aiming to improve specificity and affinity [206], in which case this approach has conceptual overlap with some tethering applications (Fig. 8.Nk) and also binding partner searches by phage display (Fig. 8.Nl). Although the assortment of the oligonucleotide tag library pairs is random, interchange will be very slow if the experimental temperature is significantly below the melting temperature of the library duplexes. With suitable common annealing sequences maintained at a temperature where duplexes are only transiently stable and equilibrate their strand interchange at a significant rate, the ESAC library would acquire characteristics of a dynamic bimolecular combinatorial library (Chapter 8 of *Searching for Molecular Solutions*). Depending on the sequence of the common annealing region (Fig. 8.Nm), DNA triplex formation is also a possibility, allowing potential trimolecular interactions with target [206].

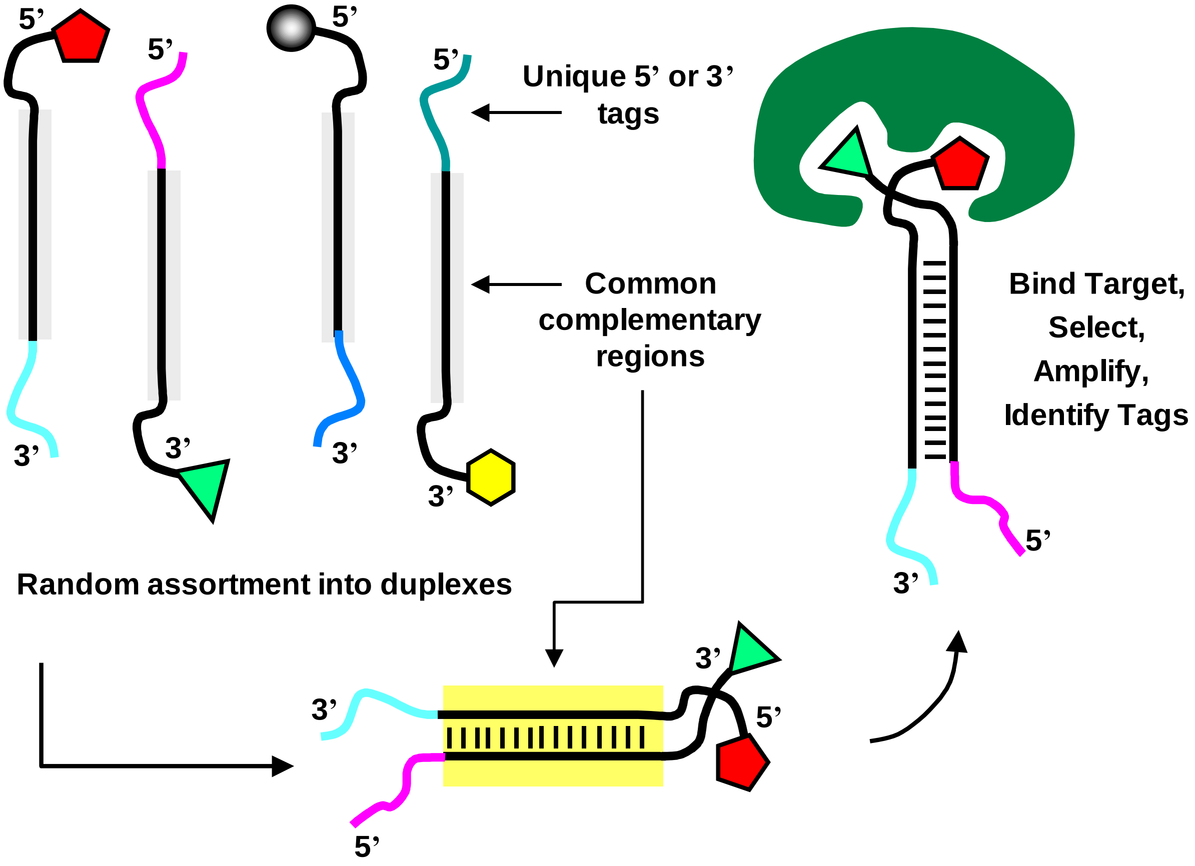

5’
5’
5’
5’
Unique 5’ or 3’ tags
Common complementary regions
3’
3’
3’
3’
Random assortment into duplexes
Bind Target, Select, Amplify, Identify Tags
3’
5’
3’
3’
5’
5’

Fig. 8.Nm

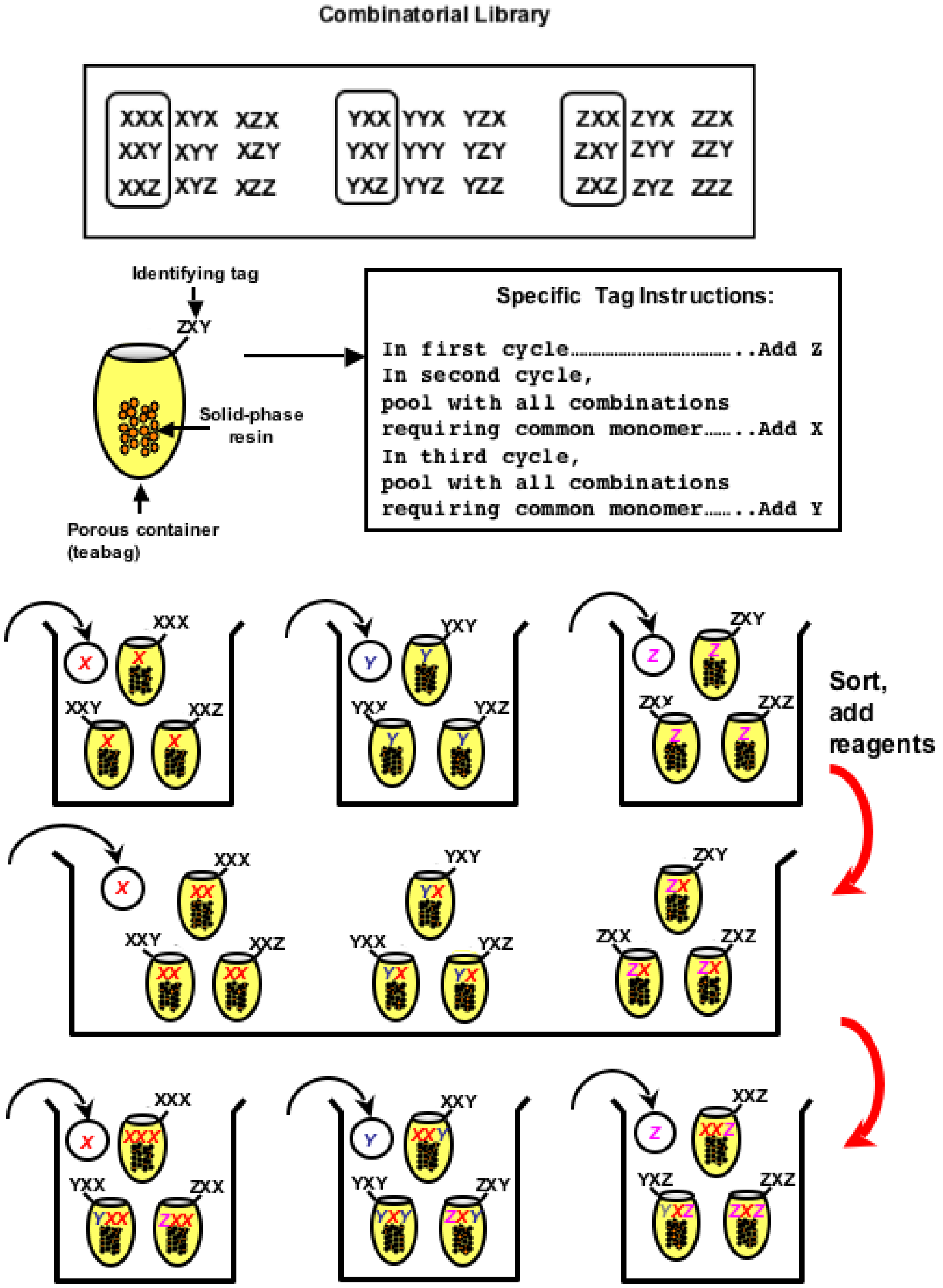
Principle of Encoded Self-Assembling Combinatorial (ESAC) libraries. A small molecule library is attached to both the 5’ and 3’ ends of oligonucleotides which bear a common region and a unique sequence tag identifier for the specific appended library member. Hybridization allows random self-assembly of the library into duplexes. The bidirectionality of the appended small molecules permits high-affinity binding to proximal target sites. Selection of binders allows their identification via their appended sequence tags.

Section 25: ***Pre-encoded Chemical Libraries***

Cited on p. 300 of *Searching for Molecular Solutions*

Combinatorial chemical libraries of limited size can be synthesized such that all sequence variations are pre-determined [207]. In other words, by successive syntheses at spatially defined sites, a combinatorial library can be built up in a 'parallel synthesis' which is 'pre-encoded' through the chosen synthetic strategy itself.

Pre-encoded combinatorial libraries in general are thus defined in both location and chemical identity during the course of synthesis and after its completion [208]. There are various ways of implementing a pre-encoded library synthesis. A useful strategy has been placing sold-phase reaction resins (on which library building blocks are assembled) into permeable containers with pores retaining the resin but permitting reactant entry (these are commonly termed 'teabags', which encapsulates their operating principle into a single evocative word). An example of a pre-encoded library synthesis with porous containers is depicted in Fig. 8.Nn below. Pre-encoding of such containers can use a color-based system with manual sorting [209] or machine-readable radiofrequency tagging with automated sorting [210,211]. Instead of enclosed resins as the solid-phase matrix, unitary polymeric gel fragments have also served for combinatorial synthesis, and cutting gels into pre-designated physical shapes can actually function as an encoding mechanism [212], although this has not been widely used.



Combinatorial Library

XXX	XYX	XZX	YXX	YYX	YZX	ZXX	ZYX	ZZX
XXY	XYY	XZY	YXY	YYY	YZY	ZXY	ZYY	ZZY
XXZ	XYZ	XZZ	YXZ	YYZ	YZZ	ZXZ	ZYZ	ZZZ

Specific Tag Instructions:

```
In first cycle……………………………..Add Z
In second cycle,
pool with all combinations
requiring common monomer……..Add X
In third cycle,
pool with all combinations
requiring common monomer……..Add Y
```

Fig. 8. Nn

Fig. 8.Nn. Encoding combinatorial synthesis by pre-encoded directed-sorting. Three monomers X, Y and Z can combine into 27 possible trimeric sequence forms (top panel). Direct-sorting synthesis of the encircled subset of these trimers is depicted in the bottom panel. Each porous container for synthesis (tea bag) is marked by an appropriate tag label (black letters) from the beginning to indicate the desired final product in each case. Containers are sorted according to the monomer requirements for the next step in the synthetic procedure. For example, in the second cycle, all members of this subset require addition of monomer X, and are thus grouped together for this step. Added monomers are color-coded as **X** (red letters), **Y** (blue letters), and **Z** (purple letters); the sequence of each product at the end of each cycle is shown with appropriate colored italic letters inside the containers, as directed by the tag shown on the outside.

In Chapter 8 of *Searching for Molecular Solutions*, the distinction between pre-encoded libraries and combinatorial chemical libraries with encoding engineered during synthesis ('in-process') was made. In the latter case, selection for a desired function also yields a chemical tag of some sort (often an amplifiable nucleic acid), which provides the information for identifying the sequence of the combinatorial library member of interest. In between complete pre-encoding of a library content and in-process informational tags, it is possible to use strategies with combinatorial syntheses which allow identification of a final active library member through the operation of iterated sequential analyses. These can either be 'recursive' (as in the following Section 26), or 'forward' with a 'positional scanning' approach. The latter strategy makes use of partially-randomized mixed combinatorial pools, conceptually related to the screening of chemical libraries by compound pooling referred to in the High-throughput Screening Section 21 above. But a significant difference in this case is that random combinatorial synthesis is used at specific positions in the oligomeric molecule of interest, while other positions are progressively fixed as the optimal residues are identified within pools of steadily decreasing complexity, until the final screening is conducted on a set of fully-defined molecular candidates. This positional scanning strategy was originated with peptides [213], and is depicted in a general 'alphabetic' sense in Fig. 8.No below. Note here, though, that spatial encoding is still used to identify pools (with partially-known sequences) giving positive signals, with iteration of the process repeated with the acquisition of increasing information, until unique species are identified. It must also be kept in mind that the use of compound pools has certain limitations, as noted earlier, and as a general rule detection of activity tends to lose sensitivity as pool size increases [214].

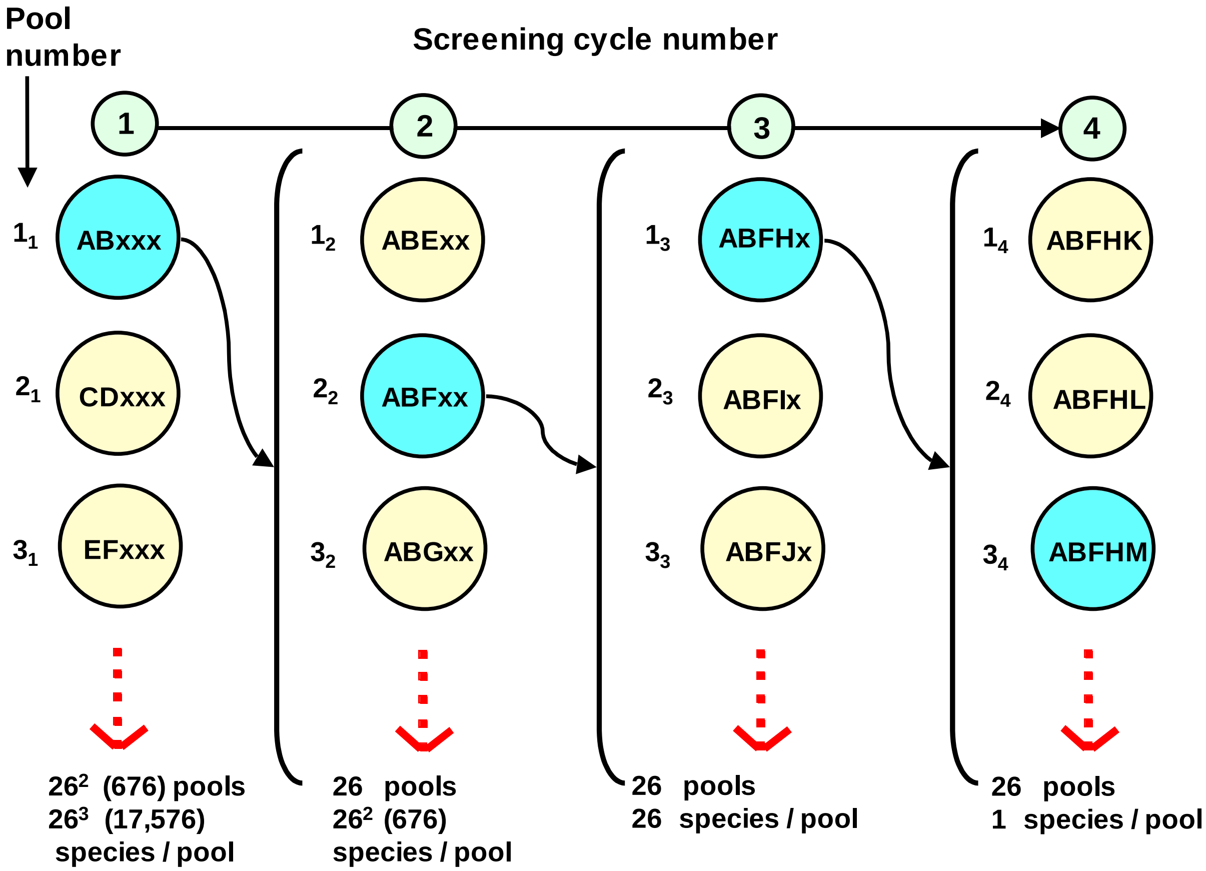


Fig. 8.No

Screening of combinatorial molecular mixtures by an iterative positional scanning procedure with progressive reduction in pool diversity until an individual molecular species is identified. In this scheme with a pentameric combinatorial molecule with five 'letters' chosen from A-Z, in the first cycle all possible combinations at the first two positions are synthesized, with random insertions in the remainder. Positive signals (aqua-colored circles) identify mixtures with optimal first-position combinations, which are then used to progressively identify optimal adjacent residues, as shown.

Section 26: ***Recursive Deconvolution***

Cited on p. 301 of *Searching for Molecular Solutions*

Topic: ***Recursive Deconvolution***

Combinatorial chemical libraries generated by mix-and-split strategies (Fig. 8.5 of *Searching for Molecular Solutions*) are amenable to analysis by a ‘proceeding backwards’ approach termed recursive deconvolution [215]. At the last step of an iterated mix-and-split synthesis, the identity of the final coupled chemical components for each combinatorial member in each split pool is known. Accordingly, screening of such pools reveals the identity of the terminal end-component in the final combinatorial pool which gives a positive signal. Now, if samples of each stage of library synthesis have been retained, the identified final end-component can be added onto each pool of the penultimate library stage. And this process is repeatable until a single active molecule is defined (Fig. 8.Np below).

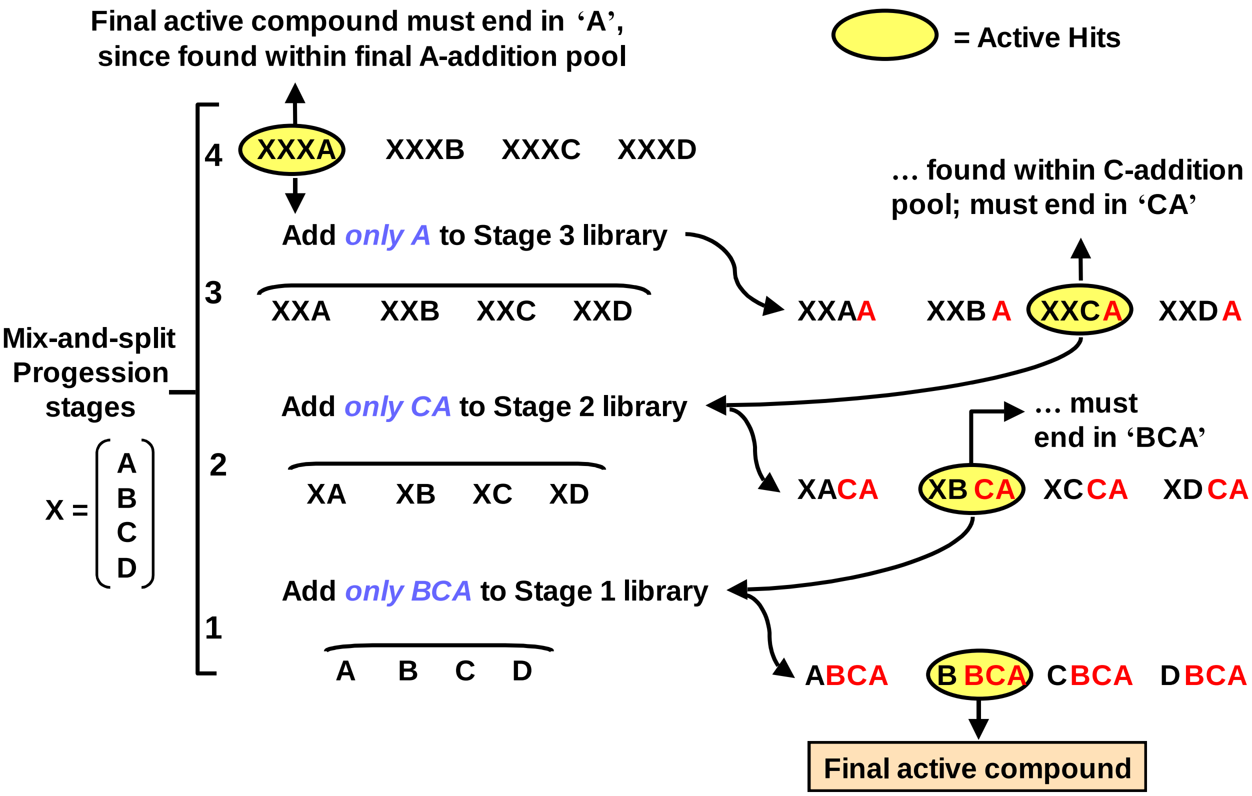

Final active compound must end in 'A', since found within final A-addition pool
= Active Hits
4
XXXA
XXXB
XXXC
XXXD
… found within C-addition pool; must end in 'CA'
Add only A to Stage 3 library
3
XXA
XXB
XXC
XXD
XXAA
XXB A
XXCA
XXD A
Mix-and-split Progession stages
Add only CA to Stage 2 library
… must end in 'BCA'
X = A B C D
2
XA
XB
XC
XD
XACA
XB CA
XCCA
XD CA
Add only BCA to Stage 1 library
1
A
B
C
D
ABCA
B BCA
CBCA
D BCA
Final active compound

Fig. 8.Np

Principle of recursive deconvolution of mix-and-split library [215]. A library with four mix-and-split stages is represented, with positive hits from screening encircled. Identification of the end-unit of the final component allows recursive addition of this to the previous-stage library, with the process repeated until an unambiguous active compound is identified.

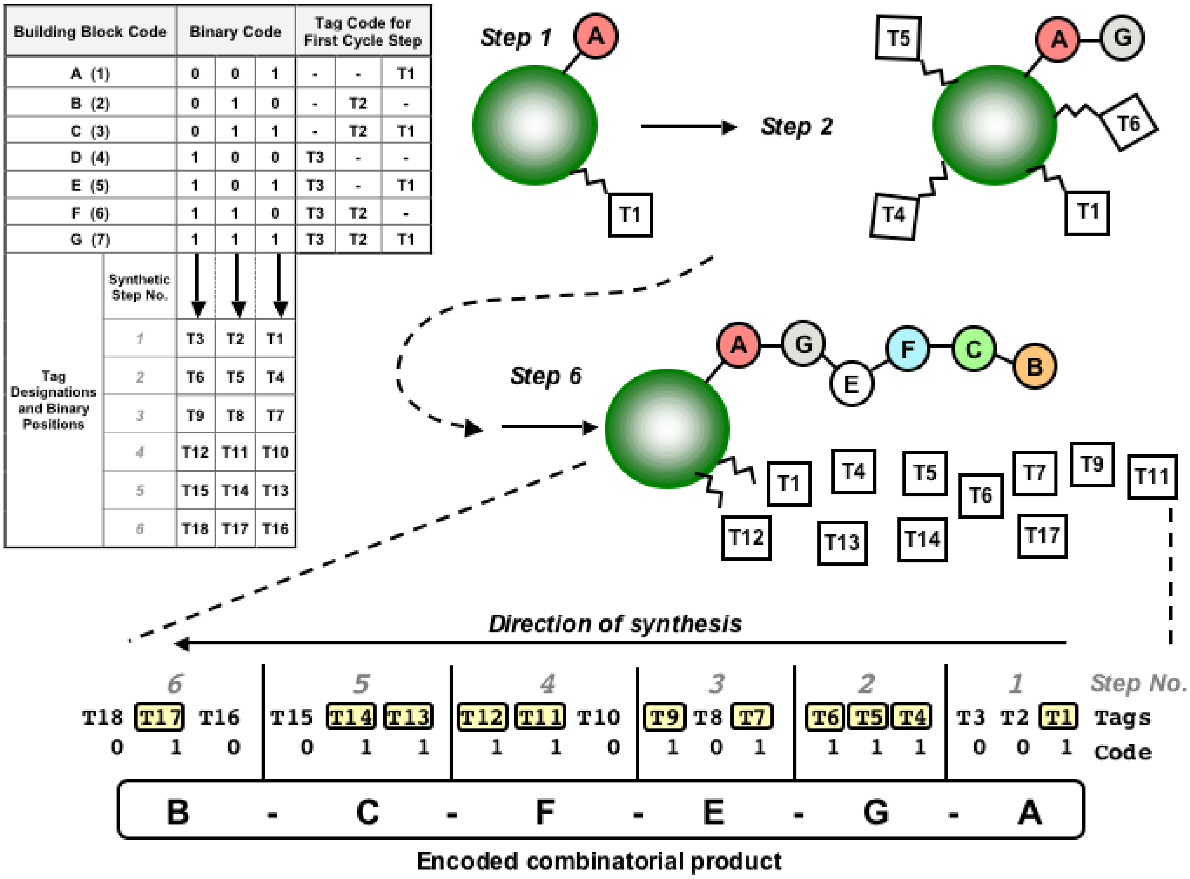
Section 27: ***Binary Codes - Libraries***

Cited on p. 304 of *Searching for Molecular Solutions*

Chemical Library Binary Code

It is not necessary to encode a sequence directly by another sequence with one-to-one coding correspondence. Codes with discrete sets of molecular tags can be designed to specify both the chemical nature of a combinatorial unit and its position in a synthetic progression. An example of this described during the relatively early stage of combinatorial chemistry development [216] is outlined schematically in Fig. 8.Nq below. Here each separate tag defines both a specific compound addition and the corresponding step number in which this addition occurred, and a synthetic product can be accordingly described with a binary code. Three distinguishable tags can code for seven building blocks of a library (for tags *T1*-*T3*, these are *T1*, *T2*, and *T3* separately, and *T1*/*T2*, *T1*/*T3*, *T2*/*T3*, and *T1*/*T2*/*T3*). If each of these building blocks is arbitrarily assigned a number from 1 to 7, and the tag numbers numerically correspond to binary positions of increasing magnitude (from right to left), then the 7 building blocks have a binary code as shown in Fig. 8.Nq (where these blocks are also given alphabetic labels A-G).

For example, block E (5th in the series) has binary number 101, which corresponds to T3 at binary third (2^2) position, blank (no T2) at binary second position (2^1), and T1 at first binary position (2^0). Thus T3 + T1 encode block E. But this can only account for the first cycle of building block couplings, so another set of tags is required to specify the second step. For six steps in total (as in Fig. 8.Nq), 18 separate tags are thus required, and each can be progressively placed into a binary series as indicated in the same figure. Screening and physical isolation of a bead giving a positive signal is followed by detachment of the tags and their analysis. In the original report, the tags were analyzed by gas chromatography, resulting in a 'bar code' read-out which enables decoding of the corresponding library combinatorial product [216]. This example also demonstrates that tags themselves can be applied in a combinatorial manner. The choice of which tags should code for which library building block is in principle arbitrary, but in practice it is important for the ease of decoding to choose tags which can be readily resolved from each other during spectrometric characterization [216].



Building Block Code	Binary Code			Tag Code for First Cycle Step		
A (1)	0	0	1	-	-	T1
B (2)	0	1	0	-	T2	-
C (3)	0	1	1	-	T2	T1
D (4)	1	0	0	T3	-	-
E (5)	1	0	1	T3	-	T1
F (6)	1	1	0	T3	T2	-
G (7)	1	1	1	T3	T2	T1

Tag Designations and Binary Positions	Synthetic Step No.			
	1	T3	T2	T1
	2	T6	T5	T4
	3	T9	T8	T7
	4	T12	T11	T10
	5	T15	T14	T13
	6	T18	T17	T16

Step 1
A
T1
Step 2
T5
A
G
T6
T4
T1
Step 6
A
G
E
F
C
B
T1
T4
T5
T6
T7
T9
T11
T12
T13
T14
T17
Direction of synthesis
6 5 4 3 2 1 Step No.
T18 T17 T16 T15 T14 T13 T12 T11 T10 T9 T8 T7 T6 T5 T4 T3 T2 T1 Tags
0 1 0 0 1 1 1 1 0 1 0 1 1 1 1 0 0 1 Code
B - C - F - E - G - A
Encoded combinatorial product

Fig. 8.Nq

Example of tagging without requiring sequencing of a code string. A binary code can be established with monomeric tags to encode both chemical combinatorial elements and synthetic positions. In this example, 7 building blocks are encoded by three tags at each coupling step, for six coupling cycles (thus requiring 18 tags in total). The binary code for the building blocks (A-G) and the corresponding tag assignment in shown in the table. Building blocks and tags are added at each synthesis step, as depicted, with all tags bound directly to the beads. After six cycles of this, an example with the above code tag collection is detached from the beads and analyzed, allowing decoding of the corresponding combinatorial library member. Tags used to encode the specific product example shown here (AGEFCB) are shown in yellow boxes in the bottom panel.

Section 28: ***DNA Display vs. CIS Display***

Cited on p. 306 of *Searching for Molecular Solutions*

The principle of biological DNA ‘CIS’ display is in the file SMS–CitedNotes-Ch4/Section 10; from the same ftp site. This is obviously quite distinct from the chemical ‘DNA display’ system for encoded small molecule coupling considered in Chapter 8 (p. 306-307) of *Searching for Molecular Solutions*)

References:

1. Kishi, Y. Complete structure of maitotoxin. *Pure Appl Chem* **70**, 339-344 (1998).
2. Nicolaou, K. C., Cole, K. P., Frederick, M. O., Aversa, R. J. & Denton, R. M. Chemical synthesis of the GHIJK ring system and further experimental support for the originally assigned structure of maitotoxin. *Angew Chem Int Ed Engl* **46**, 8875-9 (2007).
3. Nicolaou, K. C., Frederick, M. O. & Aversa, R. J. The continuing saga of the marine polyether biotoxins. *Angew Chem Int Ed Engl* **47**, 7182-225 (2008).
4. Pereira, D. A. & Williams, J. A. Origin and evolution of high throughput screening. *Br J Pharmacol* **152**, 53-61 (2007).
5. Shelat, A. A. & Guy, R. K. The interdependence between screening methods and screening libraries. *Curr Opin Chem Biol* **11**, 244-51 (2007).
6. Diller, D. J. The synergy between combinatorial chemistry and high-throughput screening. *Curr Opin Drug Discov Devel* **11**, 346-55 (2008).
7. Hunter, D. Life in the fast lane: high-throughput chemistry for lead generation and optimisation. *J Cell Biochem Suppl* **Suppl 37**, 22-7 (2001).
8. Murphy, V., Volpe, A. F., Jr. & Weinberg, W. H. High-throughput approaches to catalyst discovery. *Curr Opin Chem Biol* **7**, 427-33 (2003).
9. Meggers, E. Exploring biologically relevant chemical space with metal complexes. *Curr Opin Chem Biol* **11**, 287-92 (2007).
10. Petrascheck, M., Ye, X. & Buck, L. B. An antidepressant that extends lifespan in adult Caenorhabditis elegans. *Nature* **450**, 553-6 (2007).
11. Love, D. R., Pichler, F. B., Dodd, A., Copp, B. R. & Greenwood, D. R. Technology for high-throughput screens: the present and future using zebrafish. *Curr Opin Biotechnol* **15**, 564-71 (2004).
12. Brown, D. M., Pellecchia, M. & Ruoslahti, E. Drug identification through in vivo screening of chemical libraries. *Chembiochem* **5**, 871-5 (2004).
13. Littman, B. H. & Williams, S. A. The ultimate model organism: progress in experimental medicine. *Nat Rev Drug Discov* **4**, 631-8 (2005).
14. Dowsing, T. & Kendall, M. J. The Northwick Park tragedy--protecting healthy volunteers in future first-in-man trials. *J Clin Pharm Ther* **32**, 203-7 (2007).
15. Smith, D. A. & Schmid, E. F. Drug withdrawals and the lessons within. *Curr Opin Drug Discov Devel* **9**, 38-46 (2006).
16. Liu, B., Li, S. & Hu, J. Technological advances in high-throughput screening. *Am J Pharmacogenomics* **4**, 263-76 (2004).
17. Gillet, V. J. New directions in library design and analysis. *Curr Opin Chem Biol* **12**, 372-8 (2008).

18. Chi, K. The year of sequencing. *Nature Methods* **5**, 13-14 (2007).

19. Rusk, N. & Kiermer, V. Primer: Sequencing--the next generation. *Nat Methods* **5**, 15 (2008).

20. Strausberg, R. L., Levy, S. & Rogers, Y. H. Emerging DNA sequencing technologies for human genomic medicine. *Drug Discov Today* **13**, 569-77 (2008).

21. Aharoni, A., Griffiths, A. D. & Tawfik, D. S. High-throughput screens and selections of enzyme-encoding genes. *Curr Opin Chem Biol* **9**, 210-6 (2005).

22. Taly, V., Kelly, B. T. & Griffiths, A. D. Droplets as microreactors for high-throughput biology. *Chembiochem* **8**, 263-72 (2007).

23. Dittrich, P. S. & Manz, A. Lab-on-a-chip: microfluidics in drug discovery. *Nat Rev Drug Discov* **5**, 210-8 (2006).

24. Marcus, J. S., Anderson, W. F. & Quake, S. R. Parallel picoliter rt-PCR assays using microfluidics. *Anal Chem* **78**, 956-8 (2006).

25. Beer, N. R. et al. On-chip, real-time, single-copy polymerase chain reaction in picoliter droplets. *Anal Chem* **79**, 8471-5 (2007).

26. Laforge, F. O., Carpino, J., Rotenberg, S. A. & Mirkin, M. V. Electrochemical attosyringe. *Proc Natl Acad Sci U S A* **104**, 11895-900 (2007).

27. Melin, J. & Quake, S. R. Microfluidic large-scale integration: the evolution of design rules for biological automation. *Annu Rev Biophys Biomol Struct* **36**, 213-31 (2007).

28. Zhu, Y. & Power, B. E. Lab-on-a-chip in vitro compartmentalization technologies for protein studies. *Adv Biochem Eng Biotechnol* **110**, 81-114 (2008).

29. MacBeath, G., Koehler, A. N. & Schreiber, S. L. Printing Small Molecules as Microarrays and

Detecting Protein-Ligand Interactions en Masse. *J Am Chem Soc* **121**, 7967-7968 (1999).

30. Chiosis, G. & Brodsky, J. L. Small molecule microarrays: from proteins to mammalian cells - are we there yet? *Trends Biotechnol* **23**, 271-4 (2005).

31. Uttamchandani, M., Walsh, D. P., Yao, S. Q. & Chang, Y. T. Small molecule microarrays: recent advances and applications. *Curr Opin Chem Biol* **9**, 4-13 (2005).

32. Castel, D., Pitaval, A., Debily, M. A. & Gidrol, X. Cell microarrays in drug discovery. *Drug Discov Today* **11**, 616-22 (2006).

33. Bailey, S. N., Sabatini, D. M. & Stockwell, B. R. Microarrays of small molecules embedded in biodegradable polymers for use in mammalian cell-based screens. *Proc Natl Acad Sci U S A* **101**, 16144-9 (2004).

34. Hertzberg, R. P. & Pope, A. J. High-throughput screening: new technology for the 21st century. *Curr Opin Chem Biol* **4**, 445-51 (2000).

35. Laghaee, A., Malcolm, C., Hallam, J. & Ghazal, P. Artificial intelligence and robotics in high throughput post-genomics. *Drug Discov Today* **10**, 1253-9 (2005).

36. Parker, C. N., Shamu, C. E., Kraybill, B., Austin, C. P. & Bajorath, J. Measure, mine, model, and manipulate: the future for HTS and chemoinformatics? *Drug Discov Today* **11**, 863-5 (2006).

37. Ling, X. B. High throughput screening informatics. *Comb Chem High Throughput Screen* **11**, 249-57 (2008).

38. Zhang, J. H., Chung, T. D. & Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J Biomol Screen* **4**, 67-73 (1999).

39. Walters, W. P. & Namchuk, M. Designing screens: how to make your hits a hit. *Nat Rev Drug Discov* **2**, 259-66 (2003).

40. Silverman, L., Campbell, R. & Broach, J. R. New assay technologies for high-throughput screening. *Curr Opin Chem Biol* **2**, 397-403 (1998).

41. Eigen, M. & Rigler, R. Sorting single molecules: application to diagnostics and evolutionary biotechnology. *Proc Natl Acad Sci U S A* **91**, 5740-7 (1994).

42. Eggeling, C., Fries, J. R., Brand, L., Gunther, R. & Seidel, C. A. Monitoring conformational dynamics of a single molecule by selective fluorescence spectroscopy. *Proc Natl Acad Sci U S A* **95**, 1556-61 (1998).

43. Roeffaers, M. B. et al. Single-molecule fluorescence spectroscopy in (bio)catalysis. *Proc Natl Acad Sci U S A* **104**, 12603-9 (2007).

44. Fernandes, P. B. Technological advances in high-throughput screening. *Curr Opin Chem Biol* **2**, 597-603 (1998).

45. Klostermeier, D. & Millar, D. P. Time-resolved fluorescence resonance energy transfer: a versatile tool for the analysis of nucleic acids. *Biopolymers* **61**, 159-79 (2001).

46. Knight, A. W., Goddard, N. J., Billinton, N., Cahill, P. A. & Walmsley, R. M. Fluorescence polarization discriminates green fluorescent protein from interfering autofluorescence in a microplate assay for genotoxicity. *J Biochem Biophys Methods* **51**, 165-77 (2002).

47. Burke, T. J., Loniello, K. R., Beebe, J. A. & Ervin, K. M. Development and application of fluorescence polarization assays in drug discovery. *Comb Chem High Throughput Screen* **6**, 183-94 (2003).

48. Hafner, M. et al. Displacement of protein-bound aptamers with small molecules screened by fluorescence polarization. *Nat Protoc* **3**, 579-87 (2008).

49. Allen, M. J. et al. Identification of novel small molecule enhancers of protein production by cultured mammalian cells. *Biotechnol Bioeng* **100**, 1193-204 (2008).

50. Thomsen, W., Frazer, J. & Unett, D. Functional assays for screening GPCR targets. *Curr Opin Biotechnol* **16**, 655-65 (2005).

51. Robers, M. B., Horton, R. A., Bercher, M. R., Vogel, K. W. & Machleidt, T. High-throughput cellular assays for regulated posttranslational modifications. *Anal Biochem* **372**, 189-97 (2008).

52. Hajduk, P. J. et al. High-throughput nuclear magnetic resonance-based screening. *J Med Chem* **42**, 2315-7 (1999).

53. Hart, C. P. Finding the target after screening the phenotype. *Drug Discov Today* **10**, 513-9 (2005).

54. Eggert, U. S. & Mitchison, T. J. Small molecule screening by imaging. *Curr Opin Chem Biol* **10**, 232-7 (2006).

55. Haney, S. A., LaPan, P., Pan, J. & Zhang, J. High-content screening moves to the front of the line. *Drug Discov Today* **11**, 889-94 (2006).

56. Rausch, O. High content cellular screening. *Curr Opin Chem Biol* **10**, 316-20 (2006).

57. Kassel, D. B. Applications of high-throughput ADME in drug discovery. *Curr Opin Chem Biol* **8**, 339-45 (2004).

58. Sun, J. et al. Profiling drug membrane permeability and activity via biopartitioning chromatography. *Curr Drug Metab* **9**, 152-66 (2008).

59. Sedman, A. J. Cimetidine-drug interactions. *Am J Med* **76**, 109-14 (1984).

60. Zlokarnik, G., Grootenhuis, P. D. & Watson, J. B. High throughput P450 inhibition screens in early drug discovery. *Drug Discov Today* **10**, 1443-50 (2005).

61. Lee, M. Y. & Dordick, J. S. High-throughput human metabolism and toxicity analysis. *Curr Opin Biotechnol* **17**, 619-27 (2006).

62. Hann, M. et al. Strategic pooling of compounds for high-throughput screening. *J Chem Inf Comput Sci* **39**, 897-902 (1999).

63. Keseru, G. M. & Makara, G. M. Hit discovery and hit-to-lead approaches. *Drug Discov Today* **11**, 741-8 (2006).

64. Ganesan, A. The impact of natural products upon modern drug discovery. *Curr Opin Chem Biol* **12**, 306-17 (2008).

65. Macarron, R. Critical review of the role of HTS in drug discovery. *Drug Discov Today* **11**, 277-9 (2006).

66. Kamb, A., Wee, S. & Lengauer, C. Why is cancer drug discovery so difficult? *Nat Rev Drug Discov* **6**, 115-20 (2007).

67. Lloyd, D. G. et al. Oncology exploration: charting cancer medicinal chemistry space. *Drug Discov Today* **11**, 149-59 (2006).

68. Mayer, T. U. et al. Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. *Science* **286**, 971-4 (1999).

69. Haggarty, S. J. et al. Dissecting cellular processes using small molecules: identification of colchicine-like, taxol-like and other small molecules that perturb mitosis. *Chem Biol* **7**, 275-86 (2000).

70. von Ahsen, O. & Bomer, U. High-throughput screening for kinase inhibitors. *Chembiochem* **6**, 481-90 (2005).

71. Wilhelm, S. et al. Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nat Rev Drug Discov* **5**, 835-44 (2006).

72. Park, S. E., Min, Y. K., Ha, J. D., Kim, B. T. & Lee, W. G. Novel small molecule induces p53-dependent apoptosis in human colon cancer cells. *Biochem Biophys Res Commun* **358**, 842-7 (2007).

73. Wehner, F. et al. Indoloquinolizidine derivatives as novel and potent apoptosis inducers and cell-cycle blockers. *Chembiochem* **9**, 401-5 (2008).

74. Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W. & Friend, S. H. Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278**, 1064-8 (1997).

75. Kaelin, W. G., Jr. The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer* **5**, 689-98 (2005).

76. Tong, A. H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13 (2004).

77. Torrance, C. J., Agrawal, V., Vogelstein, B. & Kinzler, K. W. Use of isogenic human cancer cells for high-throughput screening and drug discovery. *Nat Biotechnol* **19**, 940-5 (2001).

78. Turcotte, S. et al. A molecule targeting VHL-deficient renal cell carcinoma that induces autophagy. *Cancer Cell* **14**, 90-102 (2008).

79. Dai, Y. & Grant, S. Targeting multiple arms of the apoptotic regulatory machinery. *Cancer Res* **67**, 2908-11 (2007).

80. Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924-35 (2006).

81. Thomas, J. R., Liu, X. & Hergenrother, P. J. Size-specific ligands for RNA hairpin loops. *J Am Chem Soc* **127**, 12434-5 (2005).

82. Thomas, J. R. & Hergenrother, P. J. Targeting RNA with small molecules. *Chem Rev* **108**, 1171-224 (2008).

83. Hann, M. M. & Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **8**, 255-63 (2004).

84. Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov Today* **8**, 86-96 (2003).

85. Irwin, J. J. How good is your screening library? *Curr Opin Chem Biol* **10**, 352-6 (2006).

86. Shoichet, B. K. Screening in a spirit haunted world. *Drug Discov Today* **11**, 607-15 (2006).

87. McGovern, S. L., Helfand, B. T., Feng, B. & Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J Med Chem* **46**, 4265-72 (2003).

88. McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem* **45**, 1712-22 (2002).

89. Ryan, A. J., Gray, N. M., Lowe, P. N. & Chung, C. W. Effect of detergent on "promiscuous" inhibitors. *J Med Chem* **46**, 3448-51 (2003).

90. Frenkel, Y. V. et al. Concentration and pH dependent aggregation of hydrophobic drug molecules and relevance to oral bioavailability. *J Med Chem* **48**, 1974-83 (2005).

91. Feng, B. Y. et al. Small-molecule aggregates inhibit amyloid polymerization. *Nat Chem Biol* **4**, 197-9 (2008).

92. Sanger, F. et al. The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* **125**, 225-46 (1978).

93. Fiers, W. et al. Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113-20 (1978).

94. Sutcliffe, J. G. Complete nucleotide sequence of the Escherichia coli plasmid pBR322. *Cold Spring Harb Symp Quant Biol* **43 Pt 1**, 77-90 (1979).

95. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**, 729-73 (1982).

96. Fleischmann, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).

97. Blattner, F. R. et al. The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-74 (1997).

98. Mewes, H. W. et al. Overview of the yeast genome. *Nature* **387**, 7-65 (1997).

99. C.-elegans-Sequencing-Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-8 (1998).

100. Arabidopsis-Genome-Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

101. Adams, M. D. et al. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-95 (2000).

102. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

103. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).

104. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

105. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-19 (2005).

106. Pontius, J. U. et al. Initial sequence and comparative analysis of the cat genome. *Genome Res* **17**, 1675-89 (2007).

107. Warren, W. C. et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175-83 (2008).

108. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-7 (1977).

109. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-4 (1977).

110. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat Methods* **5**, 16-8 (2008).

111. Venter, J. C., Levy, S., Stockwell, T., Remington, K. & Halpern, A. Massive parallelism, randomness and genomic advances. *Nat Genet* **33 Suppl**, 219-27 (2003).

112. Lederberg, J. E. coli K-12. *Microbiology Today* **31**, 116 (2004).

113. Tomb, J. F. et al. The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* **388**, 539-47 (1997).

114. Konturek, J. W. Discovery by Jaworski of Helicobacter pylori and its pathogenetic role in peptic ulcer, gastritis and gastric cancer. *J Physiol Pharmacol* **54 Suppl 3**, 23-41 (2003).

115. Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311-5 (1984).

116. Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. & Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, 1256-72 (2003).

117. Siew, N. & Fischer, D. Unravelling the ORFan puzzle. *Comp Funct Genom* **4**, 432-441 (2003).

118. Siew, N. & Fischer, D. Structural biology sheds light on the puzzle of genomic ORFans. *J Mol Biol* **342**, 369-73 (2004).

119. Bockaert, J. & Pin, J. P. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *Embo J* **18**, 1723-9 (1999).

120. Huang, E. S. Predicting ligands for orphan GPCRs. *Drug Discov Today* **10**, 69-73 (2005).

121. Benoit, G. et al. International Union of Pharmacology. LXVI. Orphan nuclear receptors. *Pharmacol Rev* **58**, 798-836 (2006).

122. Shi, Y. Orphan nuclear receptors in drug discovery. *Drug Discov Today* **12**, 440-5 (2007).

123. Levoye, A., Dam, J., Ayoub, M. A., Guillaume, J. L. & Jockers, R. Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers. *EMBO Rep* **7**, 1094-8 (2006).

124. Gaillard, S. et al. Receptor-selective coactivators as tools to define the biology of specific receptor-coactivator pairs. *Mol Cell* **24**, 797-803 (2006).

125. Renschler, M. F., Dower, W. J. & Levy, R. Identification of peptide ligands for the antigen binding receptor expressed on human B-cell lymphomas. *Methods Mol Biol* **87**, 209-34 (1998).

126. Williams, C. Biotechnology match making: screening orphan ligands and receptors. *Curr Opin Biotechnol* **11**, 42-6 (2000).

127. Hartley, O. The use of phage display in the study of receptors and their ligands. *J Recept Signal Transduct Res* **22**, 373-92 (2002).

128. Lespinet, O. & Labedan, B. Orphan enzymes could be an unexplored reservoir of new drug targets. *Drug Discov Today* **11**, 300-5 (2006).

129. Pouliot, Y. & Karp, P. D. A survey of orphan enzyme activities. *BMC Bioinformatics* **8**, 244 (2007).

130. Lespinet, O. & Labedan, B. Puzzling over orphan enzymes. *Cell Mol Life Sci* **63**, 517-23 (2006).

131. Robson, K. J., Chandra, T., MacGillivray, R. T. & Woo, S. L. Polysome immunoprecipitation of phenylalanine hydroxylase mRNA from rat liver and cloning of its cDNA. *Proc Natl Acad Sci U S A* **79**, 4701-5 (1982).

132. Lidsky, A. S. et al. Regional mapping of the phenylalanine hydroxylase gene and the phenylketonuria locus in the human genome. *Proc Natl Acad Sci U S A* **82**, 6221-5 (1985).

133. Kerem, B. et al. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-80 (1989).

134. Riordan, J. R. et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-73 (1989).

135. Rommens, J. M. et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059-65 (1989).

136. Weiss, K. M. Tilting at quixotic trait loci (QTL): an evolutionary perspective on genetic causation. *Genetics* **179**, 1741-56 (2008).

137. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-31 (1980).

138. Gusella, J. F. et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-8 (1983).

139. Schlotterer, C. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365-71 (2000).

140. The-International-HapMap-Consortium. The International HapMap Project. *Nature* **426**, 789-96 (2003).

141. Wagner, K. R. et al. Gentamicin treatment of Duchenne and Becker muscular dystrophy due to nonsense mutations. *Ann Neurol* **49**, 706-11 (2001).

142. Kerem, E. Pharmacologic therapy for stop mutations: how much CFTR activity is enough? *Curr Opin Pulm Med* **10**, 547-52 (2004).

143. Welch, E. M. et al. PTC124 targets genetic disorders caused by nonsense mutations. *Nature* **447**, 87-91 (2007).

144. Du, M. et al. PTC124 is an orally bioavailable compound that promotes suppression of the human CFTR-G542X nonsense allele in a CF mouse model. *Proc Natl Acad Sci U S A* **105**, 2064-9 (2008).

145. The-International-HapMap-Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

146. The-International-HapMap-Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).

147. Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).

148. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet* **19**, 135-40 (2003).

149. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**, 587-97 (2003).

150. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**, 1-14 (2001).

151. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* (2008).

152. Reich, D. E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).

153. Foster, M. W. & Sharp, R. R. Beyond race: towards a whole-genome perspective on human populations and genetic variation. *Nat Rev Genet* **5**, 790-6 (2004).

154. Conrad, D. F. et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**, 1251-60 (2006).

155. Jakobsson, M. et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003 (2008).

156. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**, 1590-605 (2008).

157. Frayling, T. M. et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-94 (2007).

158. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-5 (2007).

159. Taylor, K. D., Norris, J. M. & Rotter, J. I. Genome-wide association: which do you want first: the good news, the bad news, or the good news? *Diabetes* **56**, 2844-8 (2007).

160. Iossifov, I., Zheng, T., Baron, M., Gilliam, T. C. & Rzhetsky, A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res* **18**, 1150-62 (2008).

161. Kroymann, J. & Mitchell-Olds, T. Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435**, 95-8 (2005).

162. Iyengar, S. K. & Elston, R. C. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol Biol* **376**, 71-84 (2007).

163. Hatchwell, E. & Greally, J. M. The potential role of epigenomic dysregulation in complex human disease. *Trends Genet* **23**, 588-95 (2007).

164. van Vliet, J., Oates, N. A. & Whitelaw, E. Epigenetic mechanisms in the context of complex diseases. *Cell Mol Life Sci* **64**, 1531-8 (2007).

165. Ptak, C. & Petronis, A. Epigenetics and complex disease: from etiology to new therapeutics. *Annu Rev Pharmacol Toxicol* **48**, 257-76 (2008).

166. Wong, A. H., Gottesman, II & Petronis, A. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum Mol Genet* **14 Spec No 1**, R11-8 (2005).

167. Petronis, A. Epigenetics and twins: three variations on the theme. *Trends Genet* **22**, 347-50 (2006).

168. Weiss, K. M. Cryptic causation of human disease: reading between the (germ) lines. *Trends Genet* **21**, 82-8 (2005).

169. McCarthy, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-69 (2008).

170. Shames, D. S. et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med* **3**, e486 (2006).

171. Plass, C. & Smiraglia, D. J. Genome-wide analysis of DNA methylation changes in human malignancies. *Curr Top Microbiol Immunol* **310**, 179-98 (2006).

172. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **8**, 286-98 (2007).

173. Davies, H. et al. Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-54. (2002).

174. Vastag, B. NIH Institutes launch joint venture to map cancer genome. *J Natl Cancer Inst* **98**, 162 (2006).

175. Waltz, E. After criticism, more modest cancer genome project takes shape. *Nat Med* **12**, 259 (2006).

176. Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-8 (2007).

177. Waltz, E. Pricey cancer genome project struggles with sample shortage. *Nat Med* **13**, 391 (2007).

178. Romero, R. et al. The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. *Bjog* **113 Suppl 3**, 118-35 (2006).

179. Pennie, W. D., Tugwood, J. D., Oliver, G. J. & Kimber, I. The Principles and Practice of Toxicogenomics: Applications and Opportunities. *Toxicol Sciences* **54**, 277-283 (2000).

180. Waring, J. F. & Halbert, D. N. The promise of toxicogenomics. *Curr Opin Mol Ther* **4**, 229-35 (2002).

181. Escoubas, P., Quinton, L. & Nicholson, G. M. Venomics: unravelling the complexity of animal venoms with mass spectrometry. *J Mass Spectrom* **43**, 279-95 (2008).

182. Zartler, E. R. & Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr Opin Chem Biol* **9**, 366-70 (2005).

183. O'Connell, D. & Roblin, D. Translational research in the pharmaceutical industry: from bench to bedside. *Drug Discov Today* **11**, 833-8 (2006).

184. Vieth, M., Sutherland, J. J., Robertson, D. H. & Campbell, R. M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discov Today* **10**, 839-46 (2005).

185. Gibson, C. J. & Gruen, J. R. The human lexinome: genes of language and reading. *J Commun Disord* **41**, 409-20 (2008).

186. Weckwerth, W. & Morgenthal, K. Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* **10**, 1551-8 (2005).

187. Jacoby, E. Chemogenomics: drug discovery's panacea? *Mol Biosyst* **2**, 218-20 (2006).

188. Haggarty, S. J. The principle of complementarity: chemical versus biological space. *Curr Opin Chem Biol* **9**, 296-303 (2005).

189. Bajorath, J. Computational analysis of ligand relationships within target families. *Curr Opin Chem Biol* **12**, 352-8 (2008).

190. Gregori-Puigjane, E. & Mestres, J. Coverage and bias in chemical library design. *Curr Opin Chem Biol* **12**, 359-65 (2008).

191. Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br J Pharmacol* **152**, 5-7 (2007).

192. Frye, S. V. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. *Chem Biol* **6**, R3-7 (1999).

193. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* **5**, 262-75 (2004).

194. Savchuk, N. P., Balakin, K. V. & Tkachenko, S. E. Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr Opin Chem Biol* **8**, 412-7 (2004).

195. Miller, J. L. Recent developments in focused library design: targeting gene-families. *Curr Top Med Chem* **6**, 19-29 (2006).

196. Orry, A. J., Abagyan, R. A. & Cavasotto, C. N. Structure-based development of target-specific compound libraries. *Drug Discov Today* **11**, 261-6 (2006).

197. Kubinyi, H. Chemogenomics in drug discovery. *Ernst Schering Res Found Workshop*, 1-19 (2006).

198. Erlanson, D. A. et al. Site-directed ligand discovery. *Proc Natl Acad Sci U S A* **97**, 9367-72 (2000).

199. Erlanson, D. A., Wells, J. A. & Braisted, A. C. Tethering: fragment-based drug discovery. *Annu Rev Biophys Biomol Struct* **33**, 199-223 (2004).

200. Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Curr Opin Biotechnol* **17**, 643-52 (2006).

201. Erlanson, D. A. & Hansen, S. K. Making drugs on proteins: site-directed ligand discovery for fragment-based lead assembly. *Curr Opin Chem Biol* **8**, 399-406 (2004).

202. Erlanson, D. A. et al. In situ assembly of enzyme inhibitors using extended tethering. *Nat Biotechnol* **21**, 308-14 (2003).

203. Meyer, S. C., Shomin, C. D., Gaj, T. & Ghosh, I. Tethering small molecules to a phage display library: discovery of a selective bivalent inhibitor of protein kinase A. *J Am Chem Soc* **129**, 13812-3 (2007).

204. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* **240**, 1759-64 (1988).

205. Alber, T. Structure of the leucine zipper. *Curr Opin Genet Dev* **2**, 205-10 (1992).

206. Melkko, S., Scheuermann, J., Dumelin, C. E. & Neri, D. Encoded self-assembling chemical libraries. *Nat Biotechnol* **22**, 568-74 (2004).

207. Geysen, H. M., Meloen, R. H. & Barteling, S. J. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc Natl Acad Sci U S A* **81**, 3998-4002 (1984).

208. Affleck, R. L. Solutions for library encoding to create collections of discrete compounds. *Curr Opin Chem Biol* **5**, 257-63 (2001).

209. Guiles, J. W., Lanter, C. L. & Rivero, R. A. A Visual Tagging Process for Mix and Sort Combinatorial Chemistry. *Angew Chem Int Ed Engl* **37**, 926-928 (1998).

210. Nicolaou, K. C. Radiofrequency Encoded Combinatorial Chemistry. *Angew Chem Int Ed Engl* **34**, 2289-2291 (1995).

211. Xiao, X. Y. et al. Solid-phase combinatorial synthesis using MicroKan reactors, Rf tagging, and directed sorting. *Biotechnol Bioeng* **71**, 44-50 (2000).

212. Vaino, A. R. & Janda, K. D. Euclidean shape-encoded combinatorial chemical libraries. *Proc Natl Acad Sci U S A* **97**, 7692-6 (2000).

213. Houghten, R. A. et al. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **354**, 84-6 (1991).

214. Boger, D. L., Dechantsreiter, M. A., Ishii, T., Fink, B. E. & Hedrick, M. P. Assessment of solution-phase positional scanning libraries based on distamycin A for the discovery of new DNA binding agents. *Bioorg Med Chem* **8**, 2049-57 (2000).

215. Erb, E., Janda, K. D. & Brenner, S. Recursive deconvolution of combinatorial chemical libraries. *Proc Natl Acad Sci U S A* **91**, 11422-6 (1994).

216. Ohlmeyer, M. H. et al. Complex synthetic chemical libraries indexed with molecular tags. *Proc Natl Acad Sci U S A* **90**, 10922-6 (1993).