ELSEVIER

Chemogenomics: structuring the drug discovery process to gene families

C. John Harris and Adrian P. Stevens

BioFocus DPI, Chesterford Research Park, Saffron Walden, Essex, CB10 1XL, UK

In the post-genomic era, if all proteins in a gene family can putatively be identified, how can drug discovery effectively tackle so many novel targets that might lack structural and small-molecule inhibitory data? In response, chemogenomics, a new approach that guides drug discovery based on gene families, has been developed. By integrating all information available within a protein family (sequence, SAR data, protein structure), chemogenomics can efficiently enable cross-SAR exploitation, directed compound selection and early identification of optimum selectivity panel members. This review examines recent developments in chemogenomics technologies and illustrates their predictive capabilities with successful examples from two of the major protein families: protein kinases and G-protein-coupled receptors.

The sequencing of the human genome [1,2] has led to the identification of ~20,000–25,000 genes in human DNA [3]. This groundbreaking achievement has meant that it is now possible to identify all proteins that derive from a common gene family (more correctly, protein family). With so many new potential protein targets, drug design strategies now need to be capable of coping efficiently on the genomic scale. For example, how can novel protein family members, that might lack both structural and small-molecule inhibitor data, be explored? In response to this challenge a new field of informatics has emerged, termed chemogenomics. In this article, we review this new field and the variety of strategies that have surfaced.

Chemogenomics

Chemogenomics has been described as 'the discovery and description of all possible drugs for all possible drug targets' [4]. However, in practice it is perhaps more useful to consider chemogenomics as an approach that structures the early-stage drug discovery process in conjunction with gene (or, more correctly, protein) families. In essence, it represents a strategy to improve the efficiency of early-stage discovery through the synergistic use of all information across the target family (an excellent review of the field can be found in Ref. [5]). Chemogenomics, where analysis does not depend on knowledge of biological function, offers a complementary approach to the more traditional approach that is based on therapeutic areas, where genomically unrelated targets are addressed together. Organizing early-stage drug discovery research by gene family can significantly enhance the efficiency of drug design. It has often been observed that compounds designed and synthesized against one protein target can have useful activity against other family members. Indeed, members of the same protein family often exhibit similar biochemical and pharmacological characteristics and can share important practical aspects, such as *in vitro* assay conditions. Clearly, recognizing and reusing such similarities in early-stage drug discovery can have obvious benefits in terms of efficiency.

One area where efficiency can be improved is in the development of leads with selectivity against multiple proteins within the same gene family. By identifying all members of a gene family and classifying them into subfamilies, common elements and patterns in the sequence and tertiary structures that can be exploited in drug design are revealed. In contrast to conventional phylogenetic classifications, which utilize the entire protein sequence, chemogenomics strategies effectively redefine the exploitable biology space of the gene family, by focusing on known small-molecule binding sites. These are typically well-conserved within a gene family, and the resultant classifications are often distinct from phylogenetic ones.

Corresponding author: Stevens, A.P. (adrian.stevens@glpg.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.08.013

GLOSSARY

Common similarity metrics

A number of metrics have been developed that measure the relationship between binary strings of equal length. The following examples detail three of the most common metrics employed:

Tanimoto similarity coefficient ($Sim_T = C/(A + B - C)$): where *A* and *B* are the number of bits set to 1 in each of the bitstring descriptors, respectively, and *C* is the total number of common bits set to 1 in both descriptors. The resulting Tanimoto score is bounded between 0 and 1. If $Sim_T = 1$ then the two descriptors are identical.

Hamming distance coefficient ($Sim_H = A + B - (2C)$): where *A* and *B* are the sum of bits set in each of the bitstring descriptors, respectively, and *C* is the total number of common bits set in both descriptors. The resulting Hamming score is bounded between 0 and $(A + B)$. If $Sim_H = 0$ then the two descriptors are identical.

Cosine similarity coefficient ($Sim_C = C/(A \times B)^{1/2}$): where *A* and *B* are the number of bits set to 1 in each of the bitstring descriptors, respectively, and *C* is the total number of common bits set to 1 in both descriptors. The resulting Cosine score is bounded between 0 and 1. If $Sim_C = 1$ then the two descriptors are identical.

Conceptually, linking chemical ligand space with biological target space offers the opportunity for the translation of SAR data derived for one target to be utilized with a new target. However, the conceptual validity and, indeed, the practical success of these links are both largely dependent on the availability and quality of the protein structure and biological activity data used to define each space. Thus, chemogenomic-based strategies that attempt to predict these links have only been reported for those gene families for which there is either extensive crystal structure or, alternatively, mutagenesis data available, and where significant drug activity data have also been compiled. In the following section a selection of these predictive technologies is discussed.

Predictive chemogenomics strategies

In the past few years, several predictive chemogenomics drug-design approaches have been reported in the literature (see, for example, Refs [6–24]). The majority have been developed around a single target class, typically protein kinases (PKs), G-protein-coupled receptors (GPCRs) or nuclear hormone receptors (NHRs). These approaches tend to be tailored specifically to the target class, each exploiting (in different ways) the particular range of available small-molecule SAR information and protein structural data. However, many of the reported methodologies are not readily transferable to other target classes and are thus restricted in their chemogenomics scope. Indeed, some generalized but informative approaches, having utility across multiple target classes, have been reported. For example, in the work of Schuffenhauer *et al.* [6], on four major target classes: enzymes (e.g. PKs, proteases, etc.), GPCRs, NHRs and ligand-gated ion channels (LGICs).

To date, the majority of predictive chemogenomics strategies have focused on just two protein families: PKs and GPCRs. Vast volumes of detailed and high quality drug affinity data have been reported for both of these protein families. However, the two families differ greatly in terms of the level of structural understanding and quantity of experimental structure data that are available. In the case of PKs, in excess of 300 crystal structures of some 50 kinases have been deposited in the Protein Databank (http://www.rcsb.org), including apo forms and numerous liganded structures. By contrast, for GPCRs only one member of the mammalian family, bovine rhodopsin [25], has been structurally characterized. As a consequence, the chemogenomics strategies for each class have evolved to accommodate these differing levels of structural knowledge.

Protein kinases

PKs represent one of the largest protein families in the human genome with current estimates, in the absence of mutants and splice variants, running in excess of 500 members [26]. These enzymes play a pivotal role in intracellular signalling, gene expression regulation, cellular proliferation and cell differentiation. Protein kinsases function by catalyzing the transfer of the γ-phosphate of ATP to specific amino acid residues of protein substrates that then become activated or otherwise-modified to perform a specific cellular function [27,28]. Not surprisingly, dysfunctions of PKs are associated with numerous severe pathological states and specially designed small-molecule inhibitors have several potential therapeutic applications, notably in the areas of diabetes, immune diseases and cancer [29–33].

PKs are characterized by the presence of a conserved catalytic unit, comprising two structural lobes (Figure 1). The N-terminal lobe is composed mainly of β-sheets, whereas the C-terminal lobe

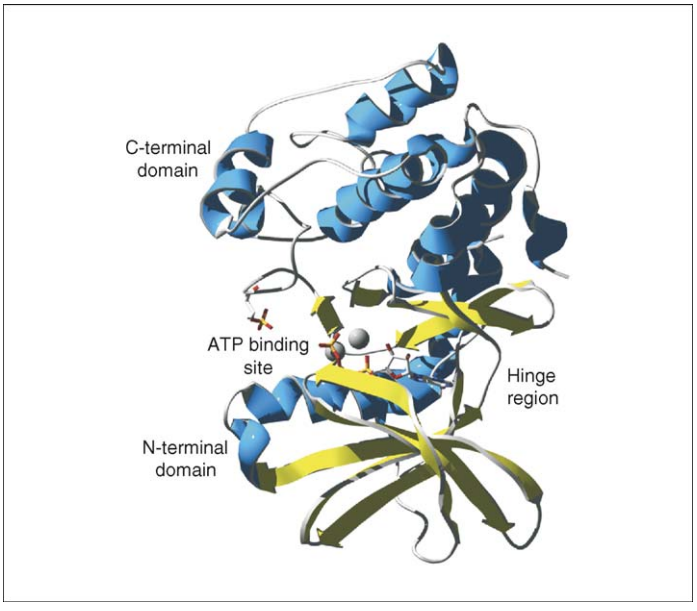


FIGURE 1

Example of the 3D structure of the protein kinase catalytic domain. The catalytic subunit is structurally well conserved among members of the superfamily, and comprises two major domains. The N-terminal domain is composed primarily of β-sheets and the C-terminal domain is dominated by α-helices. They are joined by a polypeptide strand that functions as a hinge, allowing the two domains to rotate with respect to one another. The binding site for ATP is located at the interface of these two domains and this site, together with less-conserved surrounding pockets, has been the major focus of small-molecule inhibitor design [34].

is dominated by α-helices. The two lobes are linked by a polypeptide strand that functions as a hinge, allowing the two domains to rotate with respect to one another. To date, small-molecule modulation of kinase activity has mostly been found to block ATP binding to the catalytic cleft [34], which is located between the two lobes. Although the general topology of the ATP-binding site is similar across the protein family, subtle differences in subsite conformation and electronic properties, resulting from side-chain differences, can be exploited for selective kinase drug-design purposes. As a result, several potent, selective ATP-site-directed inhibitors have been developed for a wide-range of therapeutically important kinases [34,35]; for a good introduction to PK structure, function and inhibition the reader is directed to Refs [35–37].

Ligand-centric approaches

Some of the earliest predictive chemogenomic strategies for PKs centred around the concept that an affinity profile of diverse ligands could be used to measure protein similarity [38]. Later, it was proposed that a classification of kinases based on their inhibition by ATP-competitive inhibitors would be different from groupings derived from sequence homology [39]. ter Haar *et al.* [8] used the affinity profiles of a set of 19 ligands to reclassify a diverse set of 14 PKs, and presented the results as SAR-based dendrograms. This concept was subsequently extended further by Vieth and co-workers [9,10], who reclassified 43 PKs using published selectivity data derived from >1400 inhibitory ligands. A comparison of this new SAR dendrogram with a conventional phylogenetic representation (Figure 2) illustrates that the SAR-based clusterings differ from the sequence-based clusters because of the existence of relationships between groups of targets from different subfamilies. For example, glycogen synthase kinase-3β (GSK3β) appears to be unrelated to the cyclin-dependent kinases (CDKs) in the phylogenetic classification but, by contrast, it is closely related to them in the SAR-based classifications. Each of these two representations provides a different but equally valid view of the relationships. Furthermore, they are actually complimentary and potentially enrich the medicinal chemical understanding of the protein family.

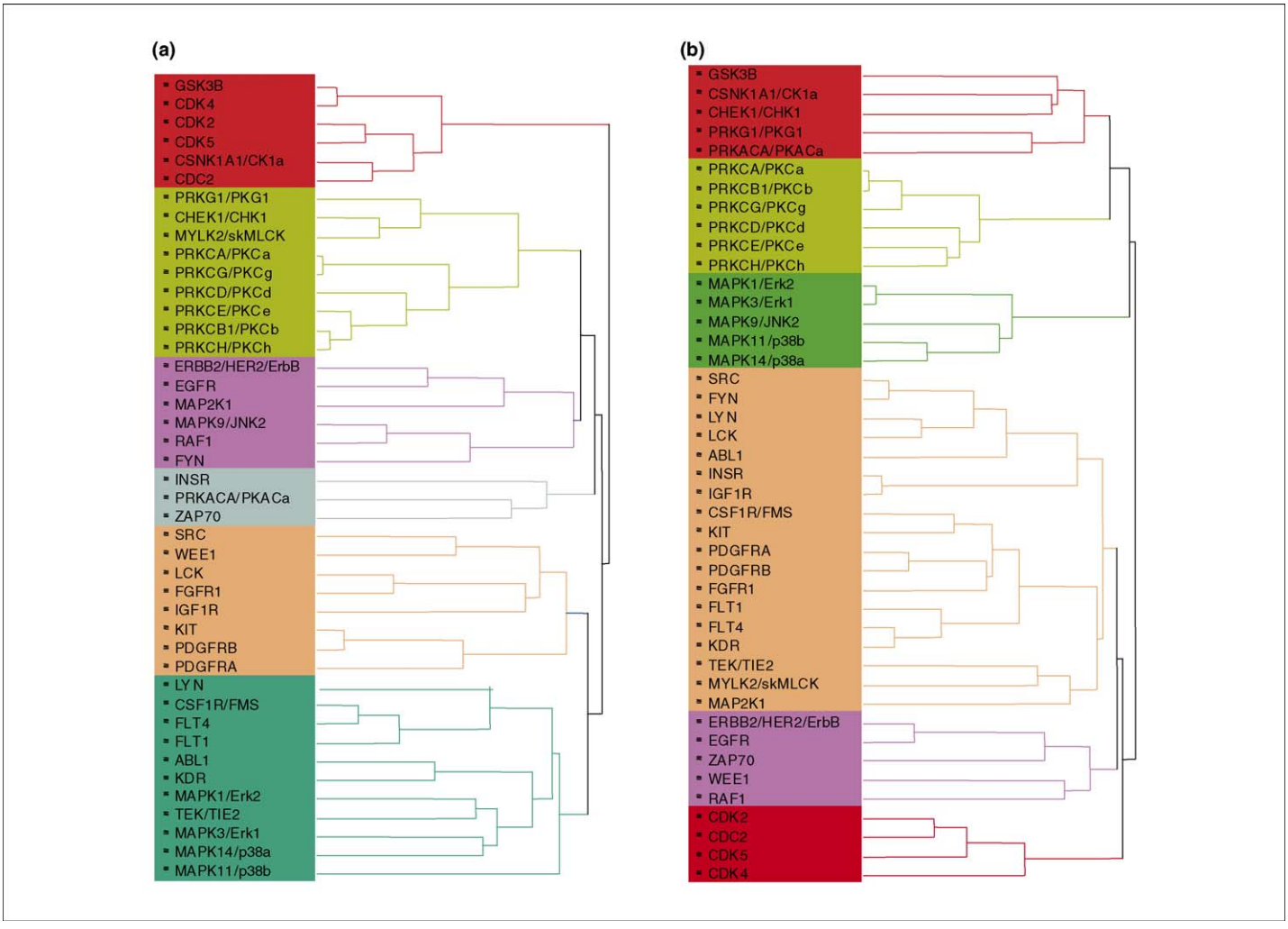


FIGURE 2

Illustration of the relative cluster-based classifications of 43 protein kinases. (a) SAR-based characterization and **(b)** conventional phylogenetic analysis using a core set of 22 amino acids from the ATP-binding site. This latter set of amino acids was determined based on a review of the interaction frequency of amino acids with small-molecule ATP-binding site inhibitors in all available X-ray structures. In each instance, protein members observed in the same cluster have been coloured equivalently. Reproduced, with permission, from Ref. [9].

Sequence-based approaches

The studies above illustrate how SAR data can be successfully used to reclassify individual members of a protein family based on their differential ability to bind a range of small-molecule inhibitors. However, this strategy is limited to those protein family members where sufficient selectivity data have been generated and it cannot be used to predict the selectivity profiles for other members of the protein family. Thus, several groups have explored the direct use of the protein sequence data as a tool for predicting protein affinity profiles.

An excellent example of such a sequence-based chemogenomic strategy for PKs is the structural interaction fingerprint (SIFt™) approach developed by Deng and colleagues [11–13]. The key to this approach is the generation of an interaction fingerprint that converts the 3D structure of the binding site into a 1D binary string. Using 34 amino acids to define the boundaries of the binding site, a bitstring descriptor (the interaction fingerprint) for each protein member is derived from the concatenation of a 7-bit interaction descriptor for each of the 34 amino acids. This SIFt™ interaction fingerprint can be treated as every other binary statistical descriptor and, thus, applied mathematically in a range of QSAR studies. For example, use of a similarity metric such as the Tanimoto score (please see Glossary) for each pairwise comparison of two kinase binding sites provides a distance matrix of similarities for every PK. This can be used in the same way as the SAR-based protein similarity studies to reclassify protein family members. However, because this approach does not make direct use of SAR data to guide the classification, the level of agreement between this and direct SAR-based strategies can vary for more-distantly related kinases.

A related sequence-based strategy has been developed by Stevens *et al.* [14]. Based on observations, from a wide-range of reported crystal structures, that the activated or partially activated conformation of the ATP-binding site when it is occupied with a bound inhibitor is broadly the same for all kinases, a generalized small-molecule binding site model was developed using a set of 31 residues (Figure 3a). Termed a 2D Roadmap™, it is applicable to all kinases within the gene family, requiring only the appropriate aligned kinase sequence to project a 2D map of key ligand-binding features. This representation of the active site was arranged so that inhibitors could also be represented in 2D, without losing implied conformational information. The model provides a rapid, simple and consistent method of representing binding-mode data for a series of kinases from *in silico* docking studies or crystallographic studies.

In common with the interaction fingerprint concept of Deng *et al.*, the 2D roadmap™ model has been used as the basis for a novel sequence-centric classification of PK similarity, termed Kinome Similarity Analysis™, (KSA™). However, a major difference between the KSA™ approach and the SIFt™ approach is its use of a weighting factor on the amino acid bitstrings. Amino acids that have side-chains (defined here as the C_α–C_β vectors) generally pointing into the ATP-binding site are more likely to participate in side-chain-based interactions with a ligand. Variations in these residues among kinases could be likely to account for differences in potency and selectivity of inhibitors. Weights were assigned to the ATP-binding site residues based on two factors: the propensity of the amino acid C_α–C_β vector to orient into the ATP-binding site; and evidence of strong correlations in amino acid type with ligand-binding SAR. Application of the Hamming metric (please see Glossary) to all pairs of amino acid bitstring descriptors gave an amino acid distance matrix from which the difference in character, attributed to any change in a particular amino acid within the ATP-binding site, could be estimated. Reduction of the matrix (which is an array of ~500 × 500 members) to, for example, two dimensions enables it to be represented graphically and, thus, yields a novel insight into the variability in molecular recognition events between family members. An example of such a plot for the ATP-binding-site view [using multidimensional scaling (MDS), a set of statistical techniques often employed in data visualization, mapping similarities and dissimilarities in matrices into lower dimensional space] is illustrated in Figure 3b. It is interesting to note that several clusters are observed that seem to be roughly equivalent to the Sugen and Hanks phylogenetic-based classifications [26,40]. This is somewhat surprising given the reduction in this model of the kinase catalytic domain from ~300 residues to ~31 residues and the associated encoding of residues.

GPCRs

Historically, GPCRs are the most commercially important class of drug targets. It is estimated that ~30% of the best-selling drugs act via modulation of GPCRs [41]. GPCRs are membrane-bound receptors that transduce extracellular signals to an intracellular response via interaction with guanine-nucleotide binding proteins (G-proteins). The range of physiological stimuli is highly diverse, ranging from photons to peptides and glycoproteins. Given their pivotal role in many intracellular signalling pathways, GPCRs continue to be of major interest for drug discovery.

Excluding olfactory receptors, it is estimated that there are ~360 druggable GPCRs, comprising three major subfamilies: class A (rhodopsin-like receptors), class B (secretin-like receptors) and class C (metabotropic glutamate-like receptors). Although a significant degree of similarity can be observed within each class, across classes GPCRs do not generally share any overall sequence homology. Despite this, GPCRs show a considerable level of structural conservation within the transmembrane domain. This domain is characterized by the presence of seven transmembrane-spanning (7TM) α-helices linked alternately by intracellular and extracellular loops (Figure 4; a good introduction to GPCR structure and function can be found in Refs [42,43]).

In contrast to PKs, the structure of only one mammalian GPCR, bovine rhodopsin [25], has been reported. Furthermore, this structure represents a GPCR in the inactivated state and is also unusual in that it has a covalently linked ligand. Because of the technical difficulties associated with obtaining the structure of membrane-bound proteins, experimental techniques, such as site-directed mutagenesis, have been used extensively to identify the domains and residues involved in ligand binding. The interpretation of mutagenesis data is not trivial, for example, in most cases it must be assumed that point mutants affect ligand binding only via direct disruption of the protein–ligand interactions, and not via indirect effects on the conformation of the GPCR. However, the general consensus is that the majority of small-molecule drugs bind to GPCRs within the 7TM domain [44]. This represents something of a fortuitous phenomenon for rational drug design given that the 7TM architecture is structurally conserved across

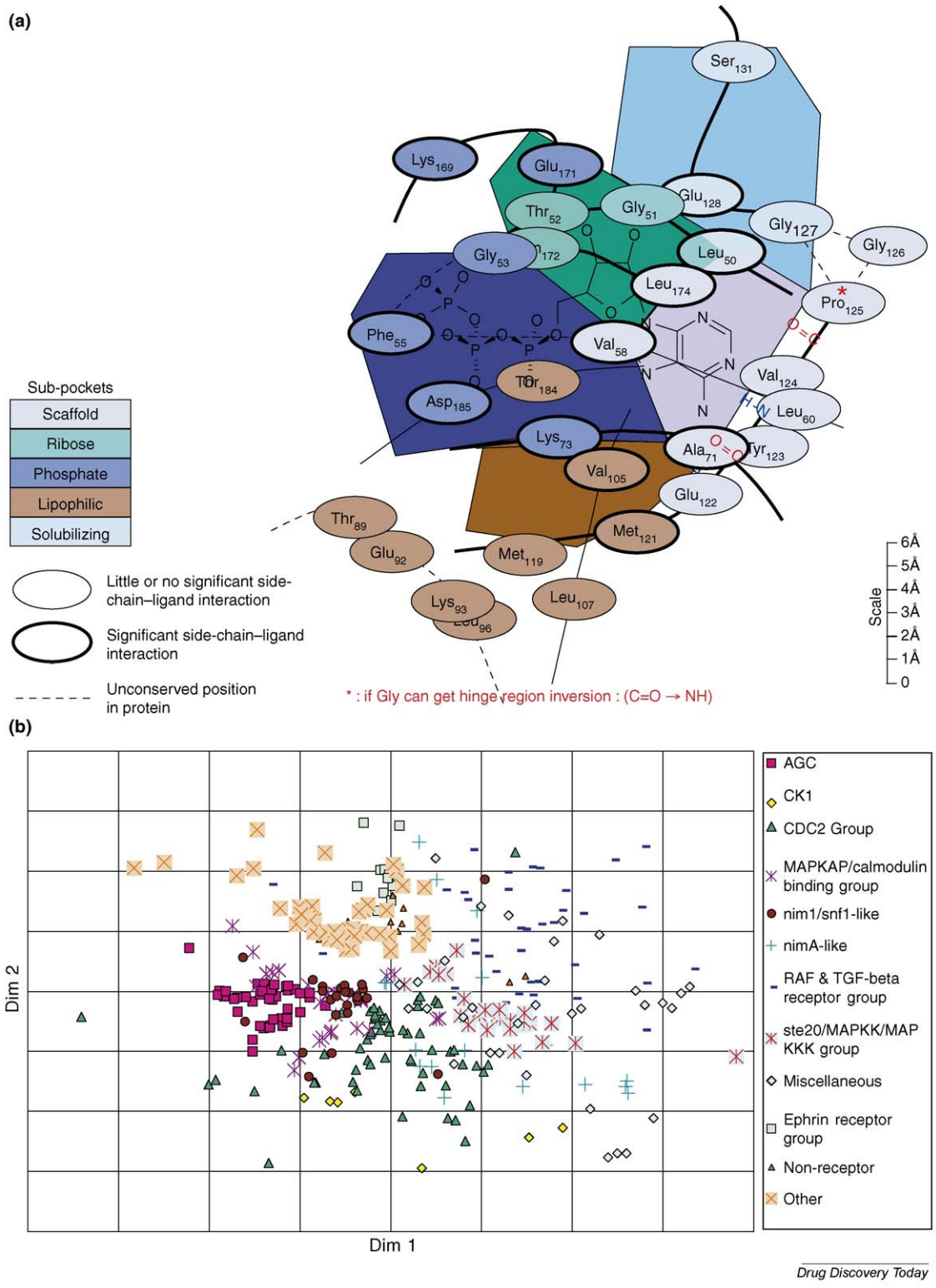


FIGURE 3

Examples of different chemogenomic-based representations of protein kinases. (a) An example of the generalized small-molecule binding site model (or 2D-Roadmap™) for cAMP-dependent protein kinase α (PKACα; KAPCA_HUMAN; accession number P17612). This model can be readily applied to all protein kinases, requiring only the appropriate aligned kinase sequence to project the 2D map of key ligand-binding features. To discriminate between ligand occupancy of different subregions of the ATP-binding site, the map is divided into subpockets using ATP as a reference ligand. On this basis, the scaffold (adenine)-binding, ribose-binding, and phosphate-binding subpockets are defined together with the two additional subpockets, labelled lipophilic and solubilizing, which are often occupied by ATP-competitive inhibitors but not ATP itself. To aid visual distinction of each of these subregions each one has been coloured differently. Furthermore, associated amino acid residues that form each subregion are coloured accordingly. **(b)** An illustration of a ligand-centric

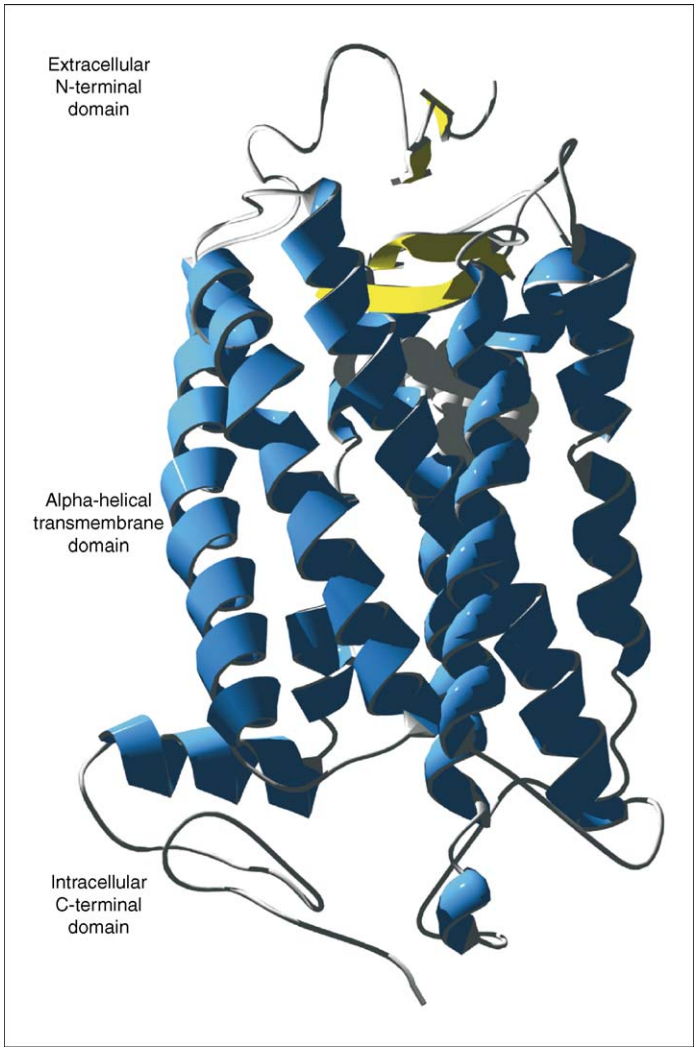


FIGURE 4

Illustration of the 3D structure of bovine rhodopsin (Protein Databank code: 1U19) that displays the characteristic seven transmembrane-spanning (7TM) α-helices linked alternately by intracellular and extracellular loops. Although bovine rhodopsin currently represents the only example of a mammalian G-protein-coupled receptor (GPCR) that has been resolved structurally [25], its 7TM architecture is considered to be representative of all GPCRs [42,43]. This represents something of a fortuitous phenomenon for rational drug design because the consensus from experimental techniques such as site-directed mutagenesis has been that the majority of small-molecule drugs bind to GPCRs within this 7TM domain [46].

the GPCR gene family, and chemogenomics approaches based on this architecture are therefore possible.

Aminergic GPCR-based chemogenomic approaches

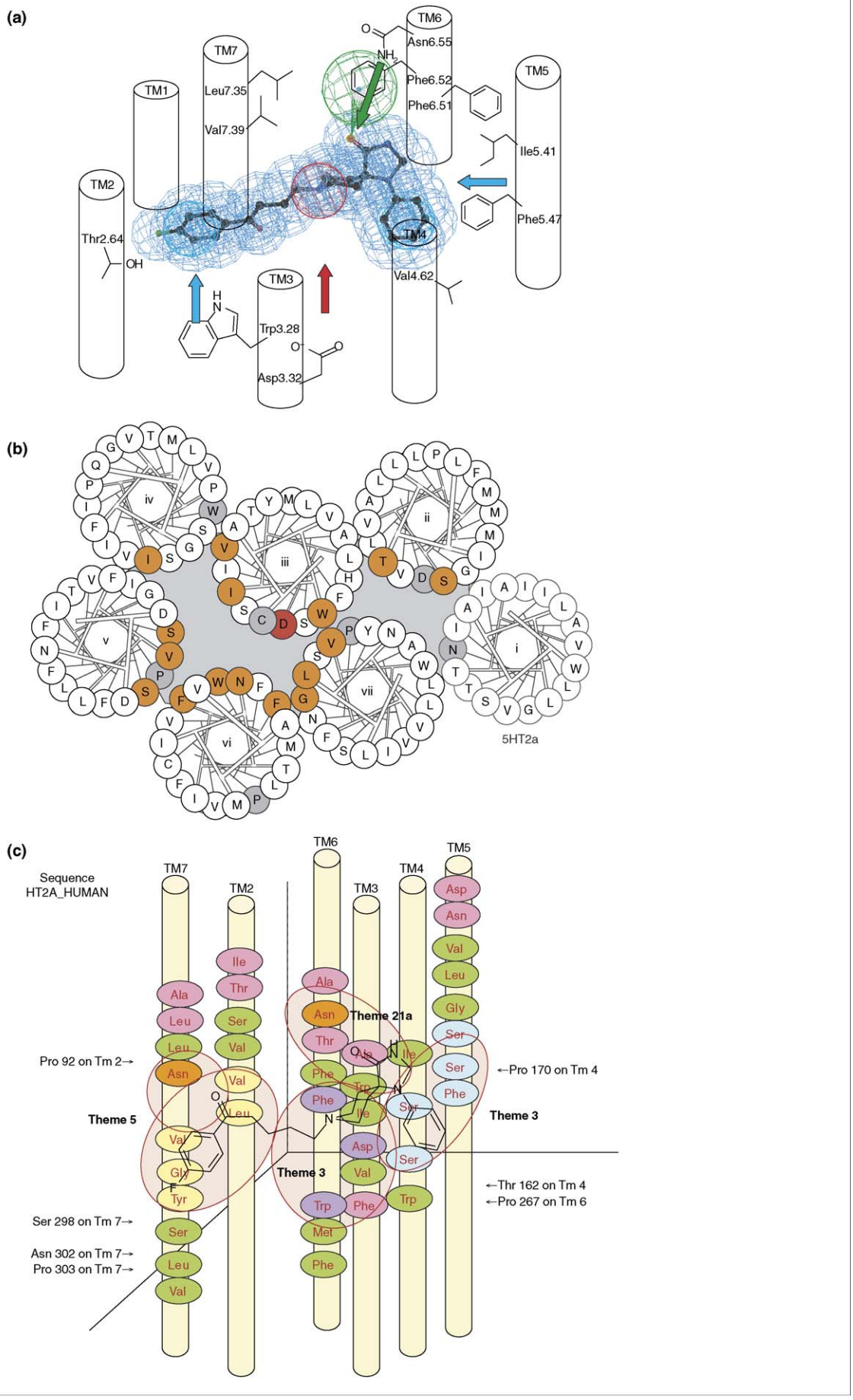
An excellent example of a GPCR chemogenomic strategy, targeting biogenic amine GPCRs, has been developed by Jacoby and colleagues [16–18]. These studies focused on examination of the small-molecule ligands that interact with the biogenic amine subset of GPCRs in relation to the amino acid residues that form the binding microenvironment within the 7TM region. This resulted in the development of a three-site binding hypothesis, consisting of the 5-hydroxytryptamine (5-HT) site, the propranolol site and the 8-hydroxy-2-di-*n*-propylamino-tetralin (8-OH-DPAT) site. These sites were subsequently further defined in terms of their sequence similarity, and the derived definitions were used to characterize a set of 50 GPCRs in terms of the presence or absence of these three sites. The significance of this approach is that GPCRs could be characterized without knowledge of any ligand. Thus, orphan receptors, whose endogenous ligands are not known, could be classified just as readily as those for which the natural ligand is known.

Another approach to the characterization of biogenic amine GPCRs has been produced by Klabunde and Evers [19]. Their primary aim was to improve the clinical safety profile of compounds progressing to development candidate status, by early identification of their putative off-target affinities. The work focused on developing models for three key GPCRs that are central in several physiological cell-signalling processes: α_{1A} adrenergic, 5-HT_{2A} serotonin and D2 dopamine receptors. The chemogenomic step in the development of these anti-target models was the derivation of a set of topographical interaction models for each receptor, to which representative drug pharmacophores were mapped by the association of putative interaction points; an example is illustrated in Figure 5a. Although the published work focused on just three common receptors, the protein features identified in these topographical maps could be readily searched and matched across other family A GPCRs.

Expanding predictive chemogenomic models to all family A GPCRs

In an approach more analogous to those of Deng and Stevens (described previously), Frimurer *et al.* [20] adopted a descriptor-based classification of family A GPCRs, termed physicogenetic analysis. Based on conventional site-directed mutagenesis and site-directed metal-binding-site engineering studies (for a description see Ref. [44]), a core set of 22 ligand-binding amino acids were identified within the 7TM domain. An empirical 5-bit bitstring was applied to encode the primary drug-recognition features of the amino acids (e.g. hydrophobic, aromatic, positive charge, negative charge and polar interactions) and the resultant distance matrix for GPCR similarities was calculated using the Cosine metric (please see Glossary). Thus, in common with both of the aforementioned PK strategies, it was subsequently possible to quantify the similarity of one receptor-binding site to other receptor-binding sites.

classification of protein kinases, using the Kinome Similarity Analysis™ (KSA™) approach [14,15]. In this approach, each amino acid that defines the bounds of the ATP-binding site is converted into a bitstring that represents its potential type of ligand contacts (e.g. charge, H-bonding, aromatic, lipophilic contact, etc.). Pairwise comparison of the bitstrings for all kinases generates a distance matrix that can subsequently be used to investigate the variation in molecular recognition between family members for small-molecule drugs. Reduction of this matrix (which is an array of ~500 × 500 members) to, for example, two dimensions using appropriate data visualization techniques (in this case, multidimensional scaling) enables it to be represented graphically. The illustrated plot for the ATP-binding-site view using multidimensional scaling reveals several clusters that seem to be roughly equivalent to the Sugen and Hanks phylogenetic-based classifications [26,42]. This is somewhat surprising, given that in this model the kinase catalytic domain is reduced from ~300 residues to ~31 residues and the associated encoding of residues.


(a)
TM6
Asn6.55
TM7
Leu7.35
Val7.39
TM1
TM5
Phe6.52
Phe6.51
Ile5.41
Phe5.47
TM2
Thr2.64
OH
TM4
Val4.62
TM3
Trp3.28
Asp3.32
(b)
5HT2a
(c)
Sequence
HT2A_HUMAN
TM7
TM2
TM6
TM3
TM4
TM5
Theme 21a
Theme 5
Theme 3
Theme 3
Pro 92 on Tm 2→
←Pro 170 on Tm 4
←Thr 162 on Tm 4
←Pro 267 on Tm 6
Ser 298 on Tm 7→
Asn 302 on Tm 7→
Pro 303 on Tm 7→

Usefully, Frimurer and colleagues also modelled the generic binding site via helical-wheel projections [45] (Figure 5b). This has the advantage that the spatial relationship between the binding site residues can be appreciated (at least at an empirical level) without the requirement to derive a homology model.

The concept of relating GPCRs via a generic small-molecule binding site model has been extended by several other groups. For example, Rognan and co-workers [21] developed a sequence-based classification approach across all major GPCR classes. Residue positions within the 7TM-binding cavity (considered relevant to the binding of inverse agonists and antagonists) were identified using the crystal structure of bovine rhodopsin. Using this approach, 369 GPCR receptors were clustered using a sequence identity score for similarity. Notably, the hierarchical clustering of the GPCRs using just 30 residues correlated well with the classical phylogenetic clustering, which is based on the full 7TM sequence. In this case, the key chemogenomic step was to exploit correlations in privileged ligand motifs with conserved sequence motifs, or 'hotspots', within the GPCRs [46]. This was illustrated in a comparative study to research reported by Bondensgaard *et al.* [46], where the structure recognition behaviour of a set of biphenyltetrazole and biphenylcarboxylic acids to a set of six GPCRs (AG22, AG2R, AG2S, GHSR, LT4R1 and LT4R2 [47–49]) was described. From visual inspection of the 30 amino acid binding site definitions of the cited GPCRs, the same set of key residues in the 7TM site were rapidly identified. Interestingly, when these residues were subsequently used as a search query against all 369 GPCRs, to discover other members that might putatively bind the same privileged structures, a further set of 17 GPCRs were identified.

A unique development of these chemogenomics strategies is enshrined in the Thematic Analysis™ approach developed by Crossley *et al.* [22–24]. Although the generalized 2D small-molecule binding site in this work is essentially similar to other GPCR models, albeit simpler, the assessment of receptor similarity has been separated conceptually from direct sequence-based similarity comparison. Instead, cliques of amino acids in the binding site have been assigned to specific themes, in accordance to their ability to match the physicochemical and electronic properties of fragments (or motifs) from associated small-molecule drugs. Once defined, each theme can be used to pattern-search the sequence of any other GPCR, which can then be classified according to the themes it contains. Notably, the approach only requires the properly aligned sequence information. Thus, it is able to model orphan receptors as readily as those for which the endogenous ligand is known. Furthermore, similar to the work of Klabunde [19] and Frimurer [20], described earlier, these data can be mapped on to a generic 3D binding site model (i.e. based on the rhodopsin structural assumption) to offer a visualization of the spatial relationship between themes and residues and motifs and entire ligands; in this way, new drug-binding hypotheses, based on previously undetected theme combinations, are developed (Figure 5c).

Application of predictive chemogenomics technologies

The various approaches described herein represent a snapshot of the chemogenomics strategies now being developed to support postgenomic drug discovery. By reorganizing SAR, sequence and protein-structure data in a manner that maximizes their value, each of these approaches can have real and practical predictive utility in drug design. Key features are the ability to:

- 'Borrow' SAR – this increases the speed of hit-to-lead programmes by exploiting SAR data from related proteins that share comparable active-site features.
- Select better screening compounds – obtain enriched subsets of compounds with increased probability of activity against a novel protein target. In comparison, *in silico* strategies typically rely on the availability of activity data for either the target or very close homologues to train selection strategies.
- Predict off-target activities – this can enable early identification of protein family members that are likely to exhibit undesired affinity for actives against the chosen target. This is of particular value for the rational selection of selectivity screen candidates, which, at present, is often based on a divergent range of nonsystematic protocols.

Conclusions

It is clear that chemogenomics approaches can have real and practical predictive value in drug design. From the earliest and rather vague use of the term, chemogenomics (or perhaps more accurately, chemoproteomics) has matured rapidly into an essential tool for drug discovery. The SAR-based and sequence-based strategies outlined in this review provide several advantages to rational drug-design programmes, but each has its limitations. Although SAR-based strategies provide excellent opportunities to relate proteins through their selectivity profiles against test compounds they are not able to offer predictions for proteins outside the training set. In comparison, sequence-based strategies can predict ligand affinity profiles for all members of a protein family. However, because they are not explicitly trained to reproduce SAR-based classifications, they lack the same level of resolution as the SAR-based approaches. Chemogenomic strategies that combine the benefits of both approaches will need to be developed.

Nevertheless, the predictive power of sequence-based methods is beginning to attract interest in the later stages of preclinical drug development, where the early identification of off-targets is so crucial, notably in recent years when several high-profile drugs have had to be withdrawn from market as a direct result of unexpected off-target activities, Vioxx® being perhaps the most prominent example. Current practice in identifying off-targets, using the kinase field as an example, is to counterscreen

FIGURE 5

Examples of different chemogenomic-based representations of the G-protein-coupled receptor, 5-hydroxytryptamine 2A (5-HT$_{2A}$; 5HT2A_HUMAN; accession number P28223). (a) The topographical interaction model. Reproduced, with permission, from Ref. [19]; **(b)** A helical-wheel projection, with key residues in the binding site highlighted. Reproduced, with permission, from Thue W. Schwartz [45] and Thomas Högberg at 7TM Pharma. **(c)** The logical map model, with key themes identified on the receptor. Essentially, each of the representations of the 5-HT2A receptor illustrated here offers equivalent information on the putative key residues important for binding small-molecule ligands.

compounds against very large enzyme and receptor panels *in vitro*. Chemogenomics has the potential, yet to be realized, to focus such counterscreening on the small number of relevant off-targets, thus saving time and the costs of such large screens. Therefore, a key challenge for predictive chemogenomics strategies in the future will be the ability to detect binding site patterns, and thus the associated SAR – not only within but across protein families. For example, several distinct protein families, such as the PKs and the heat-shock proteins (HSPs), bind ATP, but typically share little obvious structural similarities in the binding site despite their common functional ligand. It will be intriguing to see what useful patterns emerge from such distantly related systems.

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
3 She, X. *et al.* (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 431, 927–930
4 Caron, P.R. *et al.* (2001) Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* 5, 464–470
5 Kubinyi, H. and Muller, G., eds) (2004) *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*, Wiley-VCH Verlag GmbH, Weinheim
6 Schuffenhauer, A. *et al.* (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* 43, 391–405
7 Folkertsma, S. *et al.* (2005) The nuclear receptor ligand-binding domain: a family-based structure analysis. *Curr. Med. Chem.* 12, 1001–1016
8 ter Haar, E. *et al.* (2004) Kinase chemogenomics: Targeting the human kinome for target validation and drug discovery. *Mini Rev. Med. Chem.* 4, 235–253
9 Vieth, M. *et al.* (2004) Kinomics – structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* 1697, 243–257
10 Vieth, M. *et al.* (2005) Kinomics: characterising the therapeutically validated kinase space. *Drug Discov. Today* 10, 839–846
11 Deng, Z. *et al.* (2004) Structural Interaction Fingerprint (SIFt): a novel method of analysing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* 47, 337–344
12 Chuaqui, C. *et al.* (2005) Interaction profiles of protein kinases – inhibitor complexes and their application in virtual screening. *J. Med. Chem.* 48, 121–133
13 Deng, Z. *et al.* (2006) Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem.* 49, 490–500
14 Stevens, A.P. *et al.* (2004) Bringing kinases into focus. In *Proceedings of the 15th European Symposium on Structure-Activity Relationships (QSAR) and Molecular Modelling* (Aki (Şener), E. and Yalçin, I., eds), In pp. 324–326, Computer Aided Drug Design & Development Society, Turkey
15 Stevens, A. *et al.* (2006) Bringing kinases into focus: efficient drug design through the use of chemogenomic toolkits. *Curr. Med. Chem.* 13, 1735–1748
16 Jacoby, E. *et al.* (1999) A three binding site hypothesis for the interaction of ligands with monoamine G protein-coupled receptors: Implications for combinatorial ligand design. *Quant Struct.-Act. Relat.* 18, 561–572
17 Jacoby, E. (2001) A novel chemogenomics knowledge-based ligand design strategy – Application to G Protein-coupled receptors. *Quant Struct.-Act. Relat.* 20, 115–123
18 Schuffenhauer, A. and Jacoby, E. (2004) Annotating and mining the ligand-target chemogenomics knowledge space. *Drug Discov. Today: Biosilico* 2, 190–200
19 Klabunde, T. and Evers, A. (2005) GPCR Antitarget modelling: pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side effects. *ChemBioChem* 6, 876–889
20 Frimurer, T.M. *et al.* (2005) A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorg Med. Chem. Lett.* 15, 3707–3712
21 Surgand, J.-S. *et al.* (2006) A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* 62, 509–538
22 Crossley, R. *et al.* (2003) Construction of libraries of compounds focused towards receptors and other biological targets for screening and design of drugs or agrochemicals. BioFocus plc. *PCT Int. Appl. WO* 03/004147, A2
23 Crossley, R. (2004) The design of screening libraries targeted at G-protein coupled receptors. *Curr. Top. Med. Chem.* 4, 581–588
24 Crossley, R. and Slater, M.J. (2006) A reductionist approach to chemogenomics in the design of drug molecules and focused libraries. In *CHEMOGENOMICS: Knowledge-based Approaches to Drug Discovery* (Jacoby, E., ed.), pp. 85–108, World Scientific Publishing
25 Palczewski, K. *et al.* (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 289, 739–745
26 Manning, G. *et al.* (2002) The protein kinase complement of the human genome. *Science* 298, 1912–1934
27 Adams, J.A. (2001) Kinetic and catalytic mechanisms of protein kinases. *Chem. Rev.* 101, 2271–2290
28 Schenk, P.W. and Snaar-Jagalska, B.W. (1999) Signal perception and transduction: The role of protein kinases. *Biochim. Biophys. Acta* 1449, 1–24
29 Cohen, P. (2002) Protein Kinases – The major drug targets of the twenty-first century? *Nature Reviews* 1, 309–315
30 Waters, N.C. and Geyer, J.A. (2003) Cyclin-dependent protein kinases as therapeutic drug targets for antimalarial drug development. *Expert Opin. Ther. Targets* 7, 7–17
31 Druker, B.J. *et al.* (1996) Effects of a selective inhibitor of the Able tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat. Med.* 2, 561–566
32 Sawyers, C.L. (2002) Disabling Abl-perspectives on Abl kinase regulation and cancer therapeutics. *Cancer Cell* 1, 13–15
33 Kantarjian, H. *et al.* (2002) Hematoligic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukaemia. *N. Engl. J. Med.* 346, 645–652
34 Noble, M.E.M. *et al.* (2004) Protein kinase inhibitors: Insights into drug design from structure. *Science* 303, 1800–1805
35 Cherry, M. and Williams, D.H. (2004) Recent kinase and kinase inhibitor x-ray structures: Mechanisms of inhibition and selectivity insights. *Curr. Med. Chem.* 11, 663–673
36 Williams, D.H. and Mitchell, T. (2002) Latest Developments in crystallography and structure-based design of protein kinase inhibitors as drug candidates. *Curr. Opin. Pharmacol.* 2, 567–573
37 Schindler, T. *et al.* (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science* 289, 1938–1942
38 Kauvar, L.M. *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 2, 107–118
39 Frye, S.V. (1999) Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. *Chem. Biol.* 6, R3–R7
40 Hanks, S.K. and Hunter, T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* 9, 576–596
41 Klabunde, T. and Hessler, G. (2002) Drug design strategies for targeting G protein-coupled receptors. *ChemBioChem* 3, 928–944
42 Gether, U. (2000) Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr. Rev.* 21, 90–113
43 Flower, D.R. (1999) Modelling G-protein-coupled receptors for drug design. *Biochim. Biophys. Acta* 1422, 207–234
44 Kristiansen, K. (2004) Molecular mechanisms of ligand binding, signaling and regulation within G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol. Ther.* 103, 21–80
45 Schwartz, T.W. (1994) Locating ligand binding sites in 7TM receptors by protein engineering. *Curr. Opin. Biotechnol.* 5, 434–444
46 Bondensgaard, K. *et al.* (2004) Recognition of privileged structures by G protein-coupled receptors. *J. Med. Chem.* 47, 888–899
47 Ji, H. *et al.* (1994) Differential structural requirements for specific binding of nonpeptide and peptide antagonists to the AT1 angiotensin receptor. Identification of amino acid residues that determine binding of the antihypertensive drug Iosartan. *J. Biol. Chem.* 269, 16533–16536
48 Smith, R.G. *et al.* (1993) A nonpeptidyl growth hormone secretagogue. *Science* 260, 1640–1643
49 Reiter, L.A. *et al.* (1998) Trans-3-benzyl-4-hydroxy-7-chromanyl-benzoic acid derivatives as antagonists of the leukotriene B4 (LTB4) receptor. *Bioorg. Med. Chem. Lett.* 8, 1781–1786