Computational Approaches to Model Ligand Selectivity in Drug Design

Angel R. Ortiz[1,*], Paulino Gomez-Puertas[1], Alejandra Leo-Macias[1], Pedro Lopez-Romero[1], Eduardo Lopez-Viñas[1], Antonio Morreale[1], Marta Murcia[2], Kun Wang[3]

[1]*Bioinformatics Unit, Centro de Biologia Molecular "Severo Ochoa" (CSIC-UAM) Cantoblanco, 28049 Madrid (Spain);* [2]*Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021 (USA);* [3]*The Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (USA)*

Abstract: To be effective, a designed drug must discriminate successfully the macromolecular target from alternative structures present in the organism. The last few years have witnessed the emergence of different computational tools aimed to the understanding and modeling of this process at molecular level. Although still rudimentary, these methods are shaping a coherent approach to help in the design of molecules with high affinity and specificity, both in lead discovery and in lead optimization. It is the purpose of this review to illustrate the array of computational tools available to consider selectivity in the design process, to summarize the most relevant applications, and to sketch the challenges ahead.

Keywords: Bioinformatics, receptor-based drug design, docking, ligand selectivity.

THE PROBLEM AND RELEVANCE OF SELECTIVITY

To design a drug, one now normally searches for a *"magic bullet"* that binds specifically to a rationally chosen target. This reductionist approach is proving highly successful [1], however it has certain drawbacks. Probably one of the biggest is that the possibility that a designed molecule binds in places other than the target is often neglected, only to find out late in the discovery process that the candidate drug has failed. To be effective, a designed drug must discriminate successfully the macromolecular target from alternative structures present in the organism. Not only the affinity for the desired target, but also the selectivity over potential competitors, is of crucial importance. For instance, the human genome encompasses some 500 different protein-tyrosine and protein-serine/ threonine kinases [2]. A number of diseases, including cancer, diabetes, and inflammation, are known to be linked to perturbation of protein kinase–mediated cell signaling pathways, and for this reason there has been a growing interest in the use of kinase inhibitors as drugs [3]. But although it has been demonstrated exhaustively that promising ATP-antagonistic inhibitors can be found, the biochemical, cellular, and *in vivo* selectivity of such inhibitors remains unresolved. Since a great majority of these inhibitors are ATP-competitive, and protein kinases share considerably homology in their ATP-binding site, most of them do not have the level of selectivity required for a successful *in vivo* pharmacological activity [4]. This is a paradigmatic case, but the examples abound. Similar situations have been found with matrix-metalloproteinases (MMPs) [5], serine proteases [6] or nuclear receptors [7]. In all these cases, a *"design out"* of the competitors can be just as important as a *"design in"* over the target.

*Address correspondence to this author at the Bioinformatics Unit, Centro de Biologia Molecular *"Severo Ochoa"* (CSIC-UAM) Cantoblanco, 28049 Madrid, Spain; E-mail: aro@cbm.uam.es

Bioinformatics may have a primary role in addressing this problem. The spectacular advances in Genomics, Proteomics and Combinatorial Chemistry, together with the accelerating pace of biomolecular structure determination and the advent of structural genomics, are starting to provide the necessary building blocks to build successful computational strategies in order to attack the problem. Fuelled by these advances, the last few years have witnessed the emergence of different computational tools aimed to the understanding of selectivity in biomolecular systems. Nowadays, in a drug design project, it is increasingly common that the 3D structure of the target macromolecule as well as a number of its putative competitors for the drug in the genome are available or can be inferred and modeled [8]. As we shall see, this information is beginning to be incorporated using a number of tools into the ligand design process. These tools range from algorithms to discover and analyze paralogs and orthologs in genome databases or to analyze multiple sequence alignments in order to uncover family specific sequence motifs, to docking post-processing tools able to extract family specific interaction patterns from docking calculations, and to algorithms for the generation of receptor-specific scoring functions to be used in virtual screening or combinatorial library design. Although still rudimentary, altogether these methods are shaping a coherent approach to the design of molecules with high affinity and specificity, both in lead discovery and in lead optimization. It is the purpose of this review to illustrate the array of computational tools available to consider selectivity in the design process, to summarize the most relevant applications, and to sketch the challenges ahead.

COMPUTATIONAL METHODS TO MODEL AND PREDICT SELECTIVITY

Let us imagine a brave computational medicinal chemist confronted to the problem of discovering a selective molecule preferentially binding to a single member of a populated but poorly characterized protein family. How

1568-0266/06 $50.00+.00

could he/she proceed? First, our imaginary friend could ask the question of what are the most suitable sites in the protein to target. Presumably, he/she would reason, these should be functionally important sites with a differentiated shape and/or electrostatic properties between the target protein and its competitors. We will start by presenting some sequence and structure based approaches to uncover these sites. Then, our medicinal chemist would probably like to know what sort of chemical groups could bind to these selectivity sites, and how he/she could select from virtual screening searches appropriate ligands binding with these optimal groups in the expected differential way he/she seeks. We will also review some of the computational tools available for this task. Now, assuming that some of these searches actually yielded new leads, the next task of our hero is, in collaboration with fellow chemists, to optimize the lead. Consequently, he/she would expect some 3D-QSAR methods to be available for this second half of the problem. These will also be surveyed.

Sequence-Based Methods

Gene duplication and divergence over million of years of evolution have given rise to collections of related proteins, which eventually can be hierarchically classified into families and subfamilies. Highly conserved residues within each one of these groups are strong candidates to be located at functionally important sites. These residues are expected to be involved in determining the interaction specificity of the subfamily members in the binding pocket. Therefore, these residues, generally referred to as *tree determinant residues* [9] or *trace residues* [10], are the main responsible of the partition of the family into its specific functional subfamilies. Thus, the elucidation of differentiating patterns of residue conservation among the different functional subclasses can be employed to predict residues expected to be involved in functional specificity. Uncovering and targeting these residues with appropriate organic groups is an attractive strategy to design selective ligands. Two main methodological approaches are available, depending on the available information (Table **1**): supervised and unsupervised methods. In the supervised methods the discriminating classes are known in advance, and the computational method attempts to find the positions that best discriminate among functional classes given the alignment. In the unsupervised methods both classification and discriminating positions are resolved simultaneously.

Provided that a functional grouping of the proteins to discriminate pharmacologically is available, supervised methods can be applied to infer those residues conferring the observed specificity. Hannenhalli & Russell [11] proposed that key residues involved in functional specificity can be revealed by comparing specific hidden Markov models (HMM) [12] fitted to multiple sequence alignments of the proteins contained in each functional group. The comparison of the different HMMs identifies the positions in the sequence alignment that are best at discriminating between the groups modelled with the HMM, i.e. positions conserved within the subclasses, but different among them. This is achieved by computing the level of dependency of a position to a given subclass by means of the relative entropy in terms of the positional parameters defined by the HMM profile. A problem with this entropy position-dependent approach is that it only detects obvious patterns of conservation within subfamilies.

Functional specificity of proteins is believed to be more conserved among orthologs than among paralogs. Orthologs are genes in different organisms which are direct evolutionary counterparts of each other. Hence, orthologs were inherited through speciation. Paralogs, on the other hand, are genes in the same organism which evolved by gene duplication. After duplication, paralogous proteins experience weaker evolutionary pressure and their specificity diverges leading to emerging of new specificities and functions. Starting from this observation, Mirny & Gelfand [13,14] proposed a method based on mutual information formulation to identify residues which determine specificity of protein-DNA and protein-ligand recognition. The idea is to start from a family of paralogs in one genome, find orthologs for each member of the family in other genomes and then identify residues that can better discriminate between these orthologous (specificity) groups. This second part uses a statistical procedure to determine whether positions in the multiple sequence alignment (MSA) can discriminate between functional sub-families better than the sequence similarity. Since the goal is to identify residues that can discriminate between paralogous proteins (different specificity) and at the same time merge orthologs (same specificity) together, the use of mutual information as a measure of association with the specificity seems natural. However, since mutual information can be biased due to the small sample size or biased amino acid composition, it is necessary to compute the statistical significance of the mutual information which, together with the mutual information value, is finally used to predict the specificity determining residues. The authors applied the method to identify residues involved in the DNA recognition of the lacI family of bacterial regulatory proteins. Mapping of the selected residues onto a protein structure showed that most of them belong to two spatial clusters. Residues of one cluster bind the DNA, while residues of the other cluster form a ligand pocket of the protein. The result is consistent with the known function of the transcription factors in this family: they repress transcription by binding the DNA and release transcription when a particular ligand is present. A difficulty with this approach is the fact that it relies heavily on the grouping of proteins by orthology. To resolve orthology, one needs to have (almost) complete genomes of several closely related organisms. This makes the analysis significantly data demanding. Even if complete genomes are available, orthology may not be easily resolved when very similar paralogs are present or when genomes are too diverged from each other.

All previous methods need a known classification scheme. While this is the most common situation in a drug discovery project, it might also be beneficial to use unsupervised methods if, for example, function classifiers are not well defined, or putative competitors are obtained from mining genome databases, but are otherwise uncharacterized. One of the first unsupervised methods was introduced by Casari *et al*. [9] with *SequenceSpace*, based on a principal component analysis (PCA) [15] of a multiple sequence alignment. The sequence space of the protein family can be considered a multidimensional space, with as

Table 1. Summary of the Different Sequence-Based Methods to Detect Selectivity Related Residues Discussed in this Review

Sequence classifier employed		Key position identification		Refs	URL´s
		Method	Properties		
Supervised (sequence classification known)	Orthologs & Paralogs	Mutual Information	Residue mutational behavior is considered independent Binary assignment of residues to sequence groups Evolutionary time not considered in the selection Selection not necessarily involves functional residues	[13,14]	Not Available
	Biochemical Function	HMM & Entropy	Residue mutational behavior considered independent Binary assignment of residues to sequence groups Evolutionary time not considered in the selection Selection mostly involves functional residues	[11]	http://www.russell.embl.de/proust2
Unsupervised (sequence classification unknown)	Variance in Sequence Space	PCA & Cluster Analysis	Coordinated mutational behavior is modeled Residues can contribute do more than one sequence cluster Evolutionary time not considered in the selection Selection mostly involves functional residues	[9]	http://industry.ebi.ac.uk/SeqSpace/
	Evolutionary Distances	Evolutionary Trace & Gene Trees	Residue mutational behavior considered independent Binary assignment of residues to sequence groups Evolutionary time considered in the selection (Consurf) Selection mostly involves functional residues	[10,17,18, 20,21]	http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/ETserverHome.html http://consurf.tau.ac.il

many axes as residue-position variables in the multiple sequence alignment. A protein sequence is represented as a point in this space. PCA defines a new subspace in this space where most of the variability in the cloud of points is contained. The latent factors identified by PCA establish a direct link between sequence patterns (groups) and residue-position patterns, allowing the mutual assembly of the sequences in different functional subclasses, along with the cluster of the residue-positions responsible of such division (Fig. (**1**)). *Sequence Space* has been used, for example, to identify residues responsible for the differences in specificity of Ras and Ral regulatory proteins [16]. The predictions were experimentally validated showing that replacement of two specific positions produced an interchange of binding specificities.

Other methods falling in this category are the Evolutionary Trace (ET) [10] and Consurf [17] methods. Both partition a sequence-based tree at different thresholds to generate groups that may correspond to different functional subclasses. The specific patterns of sequence conservation within the groups are used to mark sub-tree discriminating residues (Fig. (**1**)). These residues are mapped onto a representative structure of the family. The original ET method, by Lichtarge *et al*. [10], obtained the optimal partition threshold subjectively by visualizing the different clusters of residues mapped onto the 3D structure. Improvements were later introduced to correct the influence of sequence redundancy [18,19], and to select an optimal threshold by determining the statistical significance of the clusters of residues mapped onto the 3D structure [20]. Armon *et al*. [17] introduced ConSurf as an improved variation of the basic scheme in ET. Instead of the UPGMA method used in ET to build the gene tree, which considers equal rates of evolution along all branches, ConSurf uses a more rigorous maximum parsimony method. ConSurf also takes into account the physicochemical properties of the amino acids. It also introduces a weighting scheme in an attempt to reduce the effect of bias in the sequence sampling. Finally, another interesting property of ConSurf is its capacity to identify the branch where an amino acid replacement takes place, circumventing the problem of establishing cutoffs to define sequence groups. Yet, ConSurf cannot account for differences in branch lengths. This problem was overcome in Rate4Site [21] by using an estimate of the rate of evolutionary mutation at each position as an indicator of the degree of the conservation. In this way, amino acid changes are weighted in terms of their branch length, an effect particularly important when very similar or very distant homologs are used as input.

A main difference between *SequenceSpace* and a tree analysis methods such ET, is the way in which both methods accomplish the division of the family into functional subclasses. Whereas the family classification arranged by a gene tree is based on the complete sequence, PCA classifies sequences based on positions weighted by their degree of conservation within the group. Positions devoid of conservation patterns do not play a significant role in the partition. This difference might confer more plasticity to PCA and can provide finer functional classifications. A second advantage of a PCA classification is that allows residue-position to belong to more than one group. Examples of the use of these methods in the analysis of several protein families are presented in the case examples section.

Mixing Sequence and Structural Information

The previously described methods make a limited use of structural information. They simply use the structure to map the discriminating residues. But if structural data is available for all proteins or it can be reliably derived, it is possible to

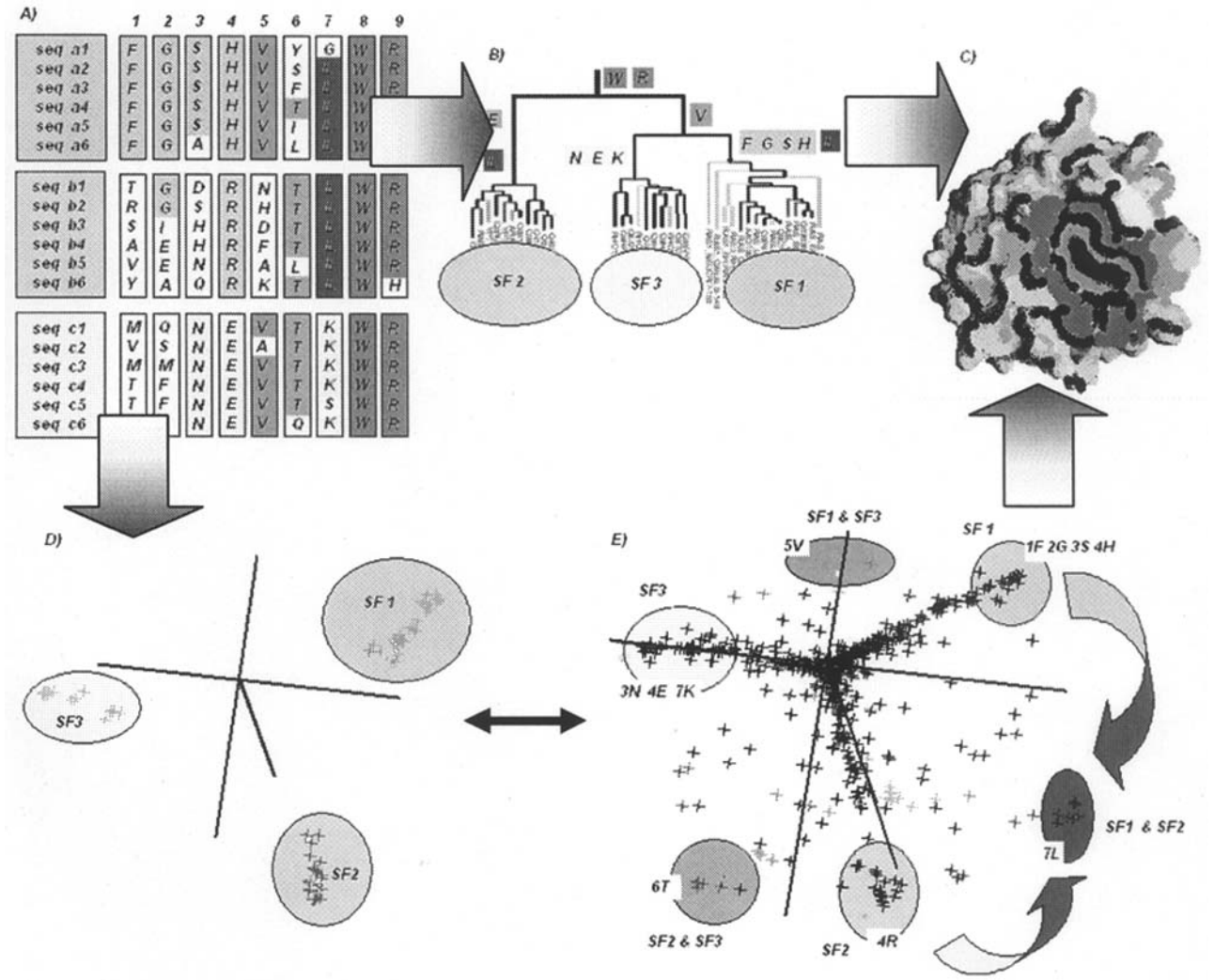

Fig. (1). Sequence Space Analysis to determine specificity residues. The analysis starts with a multiple sequence alignment (A), here representing three subfamilies, which can be obtained with programs such as ClustalW [90] or T-Coffee[91]. From the alignment it is possible to derive gene trees (B), for example with the aid of the Phylip package (http://evolution.genetics. washington.edu/phylip.html), and then assign to the different nodes in the tree those positions in the alignment holding residues clustered in a similar manner to the sequences in the subtree, using ET or ConSurf (see Table 1). These residues can be mapped onto the 3D structure of a family representative. 3D clustering might be indicative of a specificity site (C). Alternatively, the multiple sequence alignment can be subjected to Principal Components Analysis (PCA) (D). The new orthogonal axes obtained from PCA are linear combinations of the original ones. Contribution of each original variable to the determination of each axis can be inspected in plots such as the one in (E). A comparison of both plots highlights subfamilies in the alignment and specificity residues within each subfamily. For example, three subfamilies are present in (D), with subfamily three (SF3) clustering along the *y* axis. Inspection of (E) indicates that residues N, E and K at positions 3, 4 and 7, respectively, in the alignment, discriminate this subfamily from the rest. The plot also shows that T at position 6 differentiates subfamilies 2 and 3 from 1.

include this information directly in the process of determining discriminating positions. First of all, structural data can be directly incorporated in the generation of more accurate MSAs. For example, Al-Lazikani *et al.* [22] developed a strategy based on the 3D superposition of known structures and homology models combined with HMM profiling to improve the quality of the multiple sequence alignments (MSA) of the Janus Kinases (JAKs) family. The improved profiles allowed the authors to show that the human TYK2 kinase contains an SH2 domain, which should not be capable of binding phosphotyrosine, due to the presence of a histidine residue instead of the arginine residue normally found in the phosphotyrosine binding site. pKa calculations predicted that the uncharged form of histidine is predominant, likely avoiding negative charged substrates such as phosphotyrosine-containing peptides to bind.

Additionally, the spatial disposition of the residues, together with the associated physico-chemical properties of the site, such as the electrostatic potential, can also be directly used in the determination of discriminating positions. Winn *et al*. [23] explored this avenue by merging electros- tatic potentials with sequence and structural information. They used a combination of ConSurf to build a tree and to select the residues capturing the variation within

subfamilies with analysis of electrostatic potentials (ESPs) computed by solving de Poisson-Boltzmann (PB) equation [24] at different ionic strengths on homology models of the proteins in the alignment. The similarity among all pairs of ESP data was measured and a PCA study was carried out. The goal was to compare the distribution of the ESP data that of with the sequence data to reveal distinctive regions of electrostatic potential among subfamilies correlated with the sequence-based classification, and which could be associated with functional differences. They applied this method to the ubiquitin conjugating enzyme family (E2 protein family). The E2 family is known to ubiquinate histones along the ubiquitin protein ligation pathways, and have an overall negative electrostatic potential distribution. The analysis showed that an N-terminal sequence motif, likely involved in E2 binding to its E1 partner, has different ESP distributions and different functionality despite its high sequence similarity at the family level.

Similarly, Basu *et al*. [25] studied whether the ESP of purine-binding sites in proteins is sufficient to discriminate between adenine and guanine specific binding sites (A/G discrimination). The A/G discrimination at the binding site of proteins usually appears fuzzy in terms of conserved sequence and structural motifs. A PCA of the ESP distribution over the ligand-free proteins showed that the A/G binding sites can be expressed as a linear combination of two ESP principal components, representing almost completely the electrostatic component of the binding site energetics. The analysis showed that differences between the ESP patterns are enough to recognize specific from non-specific binding sites related to the A/G specificity in proteins.

The examples discussed provide strong support to the hypothesis that analysis of physico-chemical properties beyond sequence level is important to understand specificity in proteins and can be profitably used in ligand design. Novel methods based on this idea are discussed in the following paragraph.

Finding Specificity Pockets

Once key discriminating residues are selected using any of the methods previously described, selectivity pockets having strong interactions with favourably positioned organic groups in the site can be obtained. These organic groups can be incorporated in appropriate scaffolds in *de novo* design algorithms or to generate appropriate pharmacophores to search chemical libraries. Interaction grids computed around these residues can be employed to find these sites and the associated chemical groups. At each grid point interactions energies are computed between all atoms in the protein and some *typical* atoms (C, H, O, N, S, P, halogens ...) found in the majority of common drug-like molecules are computed. Equally spaced grid points surrounding the area of interest are generated, and the resultant energies stored in three-dimensional arrays for post processing. Applying PCA to these grids it is possible to extract regions discriminating groups of proteins, and therefore related to specificity. The technique, termed GRID/PCA, was first introduced by Cruciani and Goodford [26] in the field of DNA-drug interaction, and later applied by Pastor and Cruciani to the case of DHFR enzyme [27]. Kastenholz *et al*. introduced the use of multiple structures for each target to search for selectivity determinants, an approximation termed *target family landscape* [28]. The approach was successfully applied to selectivity in the case of thrombin, trypsin, and factor Xa serine protease enzymes (see below). A similar study was performed using a set of receptor structures that belong to different subfamilies of the protein kinases superfamily [29]. The analysis successfully discriminated among the different subfamilies, and highlighted key residues for selectivity (see below).

Sites selected by target family landscape or similar approaches can be targeted with virtual screening methods to detect putative binders with selective properties. Post-processing the virtual screening results, as in the *Structural Interaction Fingerprint* (SIFt) method [30], can be useful to extract these molecules from the putative binders. With SIFt, each ligand-receptor decoy is transformed in binary digits accounting for the presence (1) or absence (0) of favourable contacts between the ligands and the residues in the binding site. Each residue in each protein is coded as an N bit string consisting on 0s and 1s (N being the type of contacts to be calculated). When these strings are compared among different sequences, common interaction patterns arise, as well as key differences essentials for selectivity. If selectivity-determining residues have been spotted previously, ligands having the desired interactions with the receptors can be isolated. The method has been successfully applied to classify kinase-ligand complexes into subfamilies. It was tested for its ability as a scoring function to cluster different docking solutions, performing better than other widely used scoring functions, and as a filter in virtual screening protocols. In a recent publication from the same group, SIFt profiles (p-SIFt) were calculated by averaging bit values from the same item thorough all the sequences [31]. These profiles are used to quickly detect similarities/differences between groups of inhibitors. pSIFts have proved useful in detecting false positive and negative hits.

Determining Selectivity from First Principles

Protein-ligand binding free energy differences can in principle be computed from first principles using free energy perturbation techniques and a full atomic detail model with explicit solvent molecules using molecular dynamics simulations. However, these are computationally demanding and, therefore, cannot be used for virtual screening or library design. More affordable approaches use end-point molecular dynamics simulations and compute free energies accounting for solvent effects with continuum methods, such as MM-PBSA (Molecular Mechanics-Poisson-Boltzmann surface area) or MM-GBSA (Molecular Mechanics-Generalized Born surface area) [32]. Although still expensive, these methods can be employed to rerank a limited hit list obtained from a virtual screening run in order to select a small set of putative selective leads for experimental testing. Some recent works suggest that this might become a feasible strategy [33]. Similarly, MM-PBSA and MM-GBSA have been explored to study selectivity. For example, Rizzo *et al*. [34] performed molecular dynamics simulations with six inhibitors on stromelysin-1 and gelatinase-A, two homologous MMPs with different selectivity patterns. Selected snapshots extracted from the simulation where

processed with MM-PBSA and MM-GBSA to calculate $\Delta\Delta G_{binding}$. The calculations yielded correct values as compared with experiments. The van der Waals interactions resulted to be mainly responsible of selectivity, with a minor contribution from the electrostatic part. Aromatic rings in the binding site were shown to be the main determinants of the observed selectivity. A similar analysis was done by Laitinen *et al*. [35] who studied the binding of four 17beta-estradiols differing in the D-ring to antiestradiol antibody 57-2, antiprogesterone antibody DB3 and antitestosterone antibody 3-C4F5. Experimentally, only 17beta-estradiol is able to bind with good affinity to the antiestradiol antibody. Interestingly, a detailed analysis of the energy components allowed determining that van der Waals interactions were mainly responsible for affinity, while electrostatics accounted for the selectivity of the different ligands, in contrast to the case studied by Rizzo *et al.* [34]. In addition, $\Delta\Delta G_{binding}$ was also calculated with the computational alanine scanning method [36]. Here, each selected residue is changed, one at a time, to alanine, allowing the estimation of the side chain contribution to the binding. The higher affinity of 17beta-estradiol for the antiestradiol antibody was rationalized on the basis of hydrogen bond interactions, not present in the other steroids.

3D-QSAR Methods

So far, all methods presented here are focused on the lead generation problem. However, where selectivity issues become particularly relevant is in the lead optimization phase, when the molecular scaffold needs to be tuned to fulfill constraints other than simply binding to the target. The existing methods to estimate binding free energies using macromolecular 3D structures in docking and virtual screening cover a broad spectrum, from highly simplified scoring functions based on properties such as surface complementarity to more intensive Monte Carlo calculations involving molecular mechanics potentials. In many instances, these functions turn out to be too simplistic or insensitive to model ligand selectivity during ligand optimization. On the other hand, the more expensive free energy calculations outlined in the previous paragraph, such as those based on MM-PBSA or MM-GBSA, are too slow to efficiently explore chemical diversity.

In these cases, when activities of a representative set of chemical variations of the basic scaffold are available, it is often beneficial to resort to the use of three-dimensional quantitative structure-activity relationships (3D-QSARs). The COMBINE approach [37] is a member of this family of techniques. It is based on molecular mechanics energy computations, but it is reasonably fast, and can be easily adapted to model selectivity. The idea is that a relatively simple expression for the differences in binding affinity of a series of related ligand-receptor complexes can be derived by using multivariate statistics to correlate experimental data on binding affinities with components of the ligand-receptor interaction energy computed from the three-dimensional structures. Mathematically, the model derived from COMBINE analysis can be written as:

$$\Delta G = \sum_i w_i^{vdw} u_i^{vdw} + \sum_i w_i^{ele} u_i^{ele} + C \qquad (1)$$

where the binding affinity ΔG is given by the sum of weighted changes in energy terms, plus a constant C. The weights, w_i, are obtained from PLS analysis. In its simplest form, the one shown, just van der Waals u_i^{vdw} and electrostatic u_i^{ele} interaction energy terms are employed.

When modeling selectivity with COMBINE analysis, the chemometrics analysis is carried out with a multiple-receptor adapted version (Fig. (**2**)). A structural or sequence alignment is first used to build a position-based X-matrix of energy contributions. Lennard-Jones and electrostatics ligand-receptor interaction energies per residue are computed as usual, but introduced in a global X-matrix according to the alignment. Gaps enter the matrix with a zero value. The *y*-variable is assigned as pK_i towards the appropriate receptor.

Other 3D-QSAR methods also employed to model selectivity are CoMFA [38] (comparative molecular field analysis) and CoMSIA [39] (comparative molecular similarity indices analysis). CoMFA calculates steric and electrostatic properties according to Lennard-Jones and Coulomb potentials in a lattice around the aligned ligands. CoMSIA is an alternative approach where molecular similarity is compared in terms of similarity indices, allowing the consideration of various physicochemical properties. In both cases, differences or ratios in these properties are correlated with activity differences using multivariate statistical tools. Note that, in contrast with COMBINE, the target structure does not enter directly in the calculation of the descriptors. The resulting contribution maps are intuitively interpreted. They can also be used to map and pin down those features responsible for selectivity differences among ligands. When modeling selectivity, it is common to analyze the pairwise selectivity, using either the difference or ratio between biological activities (expressed as -log Ki) of a ligand series with respect to two different receptor types as a dependent variable. The resulting *"selectivity fields"* indicate the ways of increasing selectivity of binding, inhibition, etc… However, this type of analysis imposes limitations in multiple receptor comparisons, as they can only be obtained through multiple pairwise analyses, making the procedure cumbersome. Nevertheless, CoMFA and CoMSIA have been employed successfully to study ligand selectivity in several cases, including serine proteases [40,41], matrix metalloproteinases [42], nuclear receptors [43], glycine/NMDA and AMPA receptors [44] or protein kinases [29]. In this last case, interesting comparisons have been carried out between CoMFA and GRID/PCA methods (*vide infra*), showing good agreement with each other.

CASE EXAMPLES

Protein Kinases

This family, encoded by approximately 2% of eukaryotic genes, is one of the largest known [2]. All protein-kinases share a common domain (known as catalytic domain) of about 250-300 aminoacids, with a conserved three-dimensional structure (for a review, see Hanks [45]). These proteins are implied in a variety of cellular processes of signal transduction as cell growth, metabolism, differentiation or apoptosis. The transduction of the signal occurs

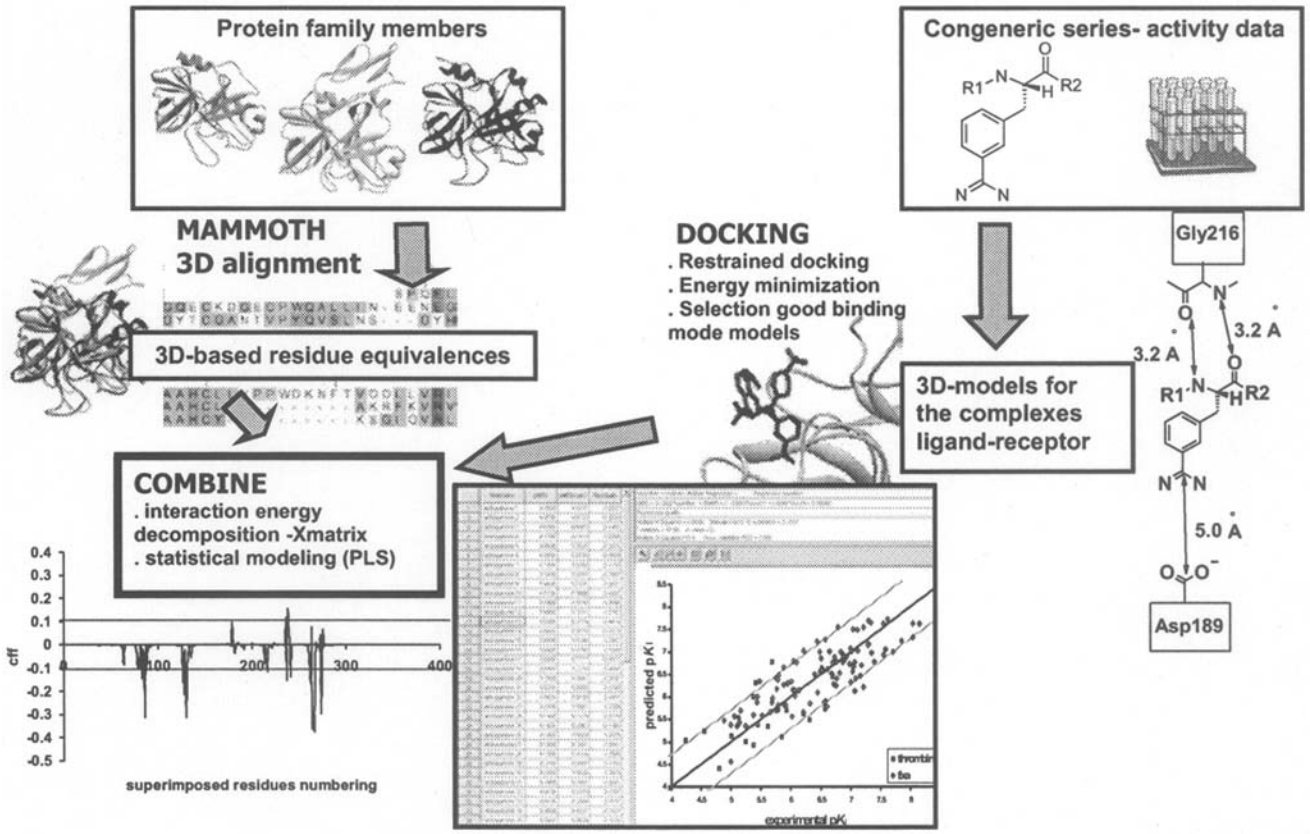

Fig. (2). Modeling selectivity with COMBINE analysis[37]. Initially, the structures in the family (either experimental or modeled structures) are structurally aligned to create a multiple structural alignment. Each ligand in the set is then docked to each one of the models. A COMBINE analysis is then carried out, using the previously computed alignment to place the interactions with the different proteins in register. A model of the predicted activity of each ligand in each protein is then generated, which can be assessed by comparison with experimental data. The coefficients of the model highlight key discriminating interactions.

through phosphorylation of specific aminoacidic residues (serine, threonine or tyrosine) in the kinase substrates, being the kinases themselves regulated by different mechanisms as phosphorylation or interaction to subfamily-specific regulatory domains. Protein kinase inhibitors are potential drugs in a number of diseases, including cancer, diabetes, inflammation or arthritis [3]. The ATP binding site is generally considered a valuable target site for binding of inhibitors, usually competitive ones. A conceptual difficulty of this approach resides in the fact that the ATP locus is mainly structurally equivalent in all known kinases, and not very different from other ATP binding proteins in the cell, which makes difficult the design of specific compounds [4] to inhibit the activity of selected kinases in precise signaling pathways without modifying the behavior of the transduction network.

Vieth *et al*. [46] performed a comparison of the classification of protein kinases, both based on sequence-based clustering, and on small molecule selectivity information, using published experimental data. The authors investigated in particular the kinase gatekeeper hypothesis: the residue identified to be one of the major determinants for selectivity in MAP kinases [47,48]. The similarity of kinase selectivity profiles was compared with the presence of small (T, S, A, V, C) or large aminoacids in this precise position for both, or only one, proteins in the pair. While kinases with similar sizes in their gatekeepers showed comparable profiles, the available information was shown to be insuffi-cient to explain numerically the inhibitor selectivity data. High sequence or structural similarity implied similar trends in ligand recognition, but the lack of sequence identity was not necessarily related to differences in inhibitor selectivity.

On the other hand, Naumann & Matter [29], using a set of 26 X-ray structures of different protein kinases, applied *"target family landscape analysis"* to find subfamily-specific interaction patterns in the ligand binding site. The first principal component in the PCA analysis was able to separate cyclin-dependent kinases (CDKs) and MAP kinases from the big family of cAMP dependent kinases, while the second principal component allowed differentiating MAP kinases from CDKs. To complement their findings, the authors performed a comparison of experimental affinities to CDK1 with affinity predictions based on final 3D-QSAR models of comparative molecular field analysis (CoMFA) [38] and comparative molecular similarity index analysis (CoMSIA) [39] for a series of 86 2,6,9-substituted purines as CDK inhibitors, including the potent purvalanol B. Consistent results were obtained.

p-SIFt has been applied [31] as post processing and re-scoring tool of the 30 top solutions obtained using the flexible docking algorithm FlexX [49] on a set of 93 X-ray structures of 21 different protein kinases complexed with ATP or ATP analogues and a variety of small molecule ligands. The initial SIFt analysis resulted in the clustering of the structures in three groups: p38 cluster (nine structures containing different inhibitors interacting with p38 kinase), CDK2 cluster (20 structures including CDK2 inhibitors) and ATPg cluster (nine and sixteen structures containing ATP or ATP analogs respectively bound to different kinases) remaining 49 structures unclassified. The obtained clusters were used to generate profiles to represent the degree of interaction conservation, defining a p-SIFt from the SIFts of those structures that were subsequently used to scoring the ligand selectivity among clusters. The results were variable, needed for a better definition of the p-SIFt in the case of the small p38 cluster, although the method appeared to discriminate correctly in the case of the larger CDK2 and ATPg clusters.

Carnitine/Choline Acyltransferases

Based on their enzymatic activity, four subfamilies of carnitine/choline acyltransferases can be differentiated (Fig. (**3**)): the carnitine palmitoyltransferases (CPTs), subdivided in CPT I and CPT II, are essential for mitochondrial β-oxidation and are located in the outer and inner mitochondrial membrane, respectively. CPT I facilitates the transfer of long-chain fatty acids from the cytoplasm to the mitochondrial matrix, which is the rate-limiting step in β-oxidation [50]. Mammalian tissues express three isoforms of CPT I, in liver (L-CPT I), muscle (M-CPT I) and brain (CPTI-C) [51-53]. Carnitine octanoyltransferase (COT) facilitates the transport of medium-chain fatty acids from peroxisomes to mitochondria through the conversion of acyl-CoAs into acylcarnitine [54]. Carnitine acetyltransferase (CrAT) catalyzes the reversible conversion of acetyl-CoA and carnitine to acetylcarnitine and free CoA [55,56]. Choline acetyltransferase (ChAT) catalyzes a similar reaction to CrAT, with the exception that the acetyl group from acetyl-CoA is transferred to choline rather than carnitine [57]. In addition to these differences in substrate recognition, the activity of the enzyme groups L-CPT I, M-CPT I and COT, but not CPT II, CrAT and ChAT, is regulated by the physiologic inhibitor malonyl-CoA [50], being the most important regulatory step in mitochondrial fatty acid oxidation. Understanding Malonyl-CoA selectivity could be valuable for the treatment of disorders such as diabetes, insulin resistance, obesity and coronary heart disease.

Efforts have been made in the analysis of the structural discrimination mechanism used by malonyl-CoA. In 2003, Morillas *et al*. [58] performed an exhaustive analysis of residues shared by all malonyl-CoA-regulated enzymes *vs.* the malonyl-CoA nonregulated members of the family, using the *SequenceSpace* algorithm [9,59,60]. Using the clusters defined by the axes corresponding to dimensions two and four, the authors found a group of five residues shared by all proteins sensitive to malonyl-CoA (L-CPT I, M-CPT I and COT), but different in the rest of the family sequences. One of such residues, the Met593 in human L-CPT I, was revealed as key for discriminating both groups of enzymes. In fact, when a mutant of L-CPT I was generated by replacing Met593 by Ser (the residue present in the non regulated group), the sensitivity of the enzyme toward malonyl-CoA was practically abolished (IC_{50} of 258 μM *vs.* 12.3 μM of the wild type). The position of Met593 in the 3D structure of a model of L-CPT I localizes the residues in the proximity of the substrate channel, explaining at least partially the experimental results (Fig. **3**). The mutant "L-CPT I M593S" has been revealed as a powerful tool for the study and perhaps treatment of Carnitine palmitoyl-transferase activity-related diseases, as diabetes, where the malonyl-CoA/CPTI interaction could be a critical step of the metabolic signaling network that controls insulin secretion [61].

The question related to the recognition of acyl-CoA substrates with different lengths of acyl chain (short–acetyl: CrAT and ChAT; medium-octanoyl: COT; large-palmitoyl: L-CPT I, M-CPT I, CPT I-C and CPT II) was also analyzed by the same authors [62-65], again with the aid of *SequenceSpace*. The most interesting finding was the discovery of a deep pocket, defined by the secondary structure elements alpha helix 12 and beta strands 1, 13, and 14, which opens to the main substrate channel, likely involved in the allocation of the long or medium acyl chains in the CPTs or COT enzymes. Although the same pocket is present in CrAT and ChAT enzymes, the entrance in these two later proteins is blocked by the side chains of metionine or cysteine residues, allowing only the entry of small acetyl groups. In contrast, the same position is occupied by glycin in COT and CPTs, allowing docking of larger substrates. A CrAT mutant with Met564 replaced by glycin confirmed the prediction: the activity of the M564G-CrAT mutant toward long chain acyl-CoAs was 1250-fold higher than that of the wild-type CrAT. In the reverse case, replacement of the equivalent Gly553 in COT by metionine (G553M-COT) resulted in markedly decreased activity toward medium and long chain natural substrates and increased activity toward short-chain acyl-CoAs [65].

Serine-Proteases

Trypsin-like serine proteases are a large family of enzymes involved in the hydrolysis of peptide bonds. Although they all act *via* a similar catalytic mechanism, they have different preferences for the amino acids that they prefer to cleave. Serine proteases play a key role in a diversity of diseases [66]. Modest changes in sequence and shape of their substrate binding sites confer to this class of enzymes a wide variety of biological functions. For instance, thrombin and fXa are prominent players in the blood clotting cascade, while trypsin is an enzyme excreted by the pancreas to aid in the digestion of nutrients. There is considerable interest in the design of selective inhibitors of these enzymes, to minimize side effects of thrombin/fXa inhibitors and to enhance their bioavailability.

It is well known that substrate selectivity in serine-proteases is conferred by key changes in a specificity pocket. Three positions were proposed originally to define the pocket. An aspartic acid found in trypsin (Asp189) is usually replaced by a small residue in chymotrypsins (Ser) and

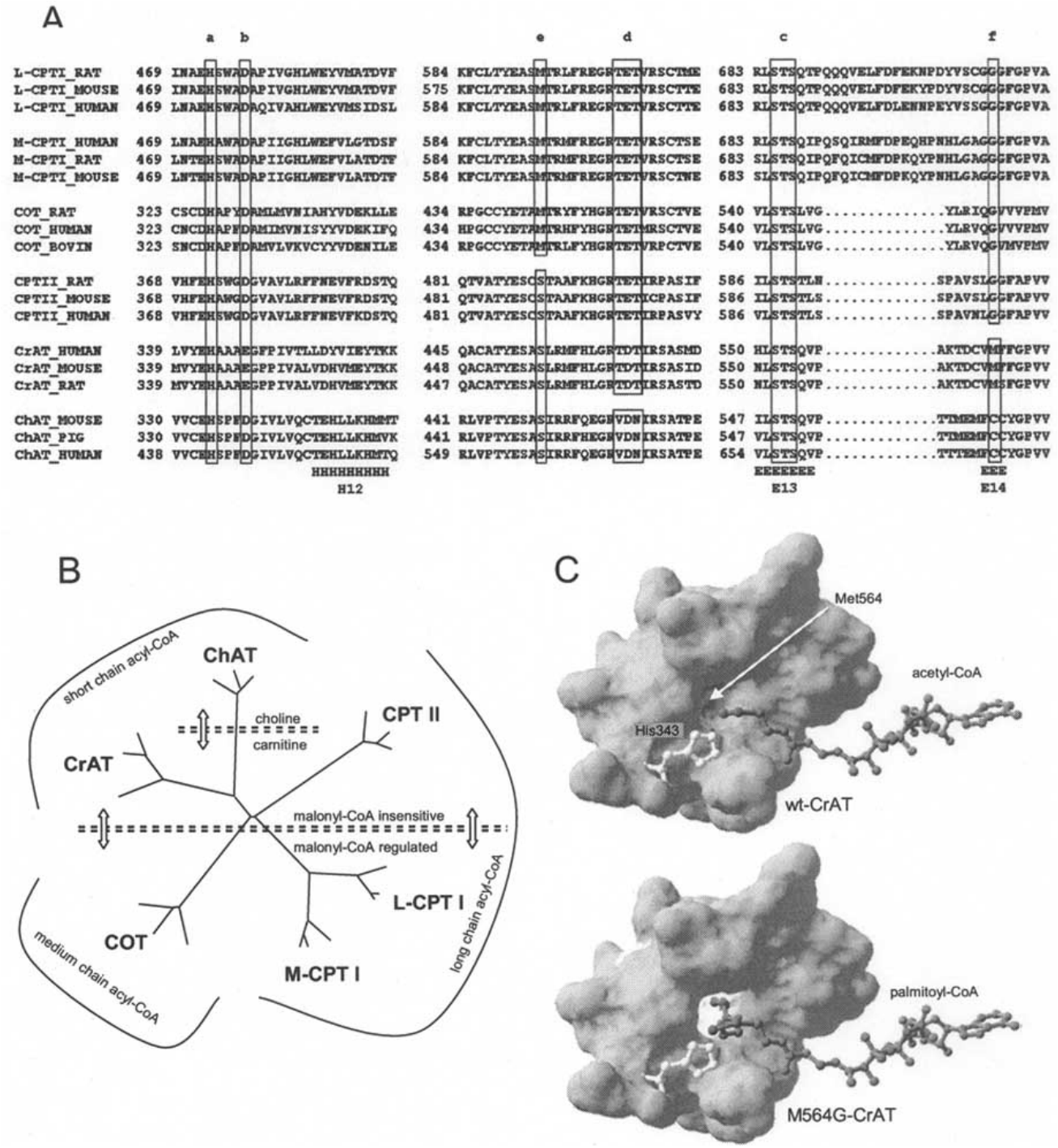

Fig. (3). Sequence and structure determinants of choline/carnitine acyltransferases. A. MSA of different subfamily representatives: carnitine palmitoyltransferase I, isoforms L (L-CPTI) and M (M-CPTI), carnitine octanoyltransferase (COT), carnitine palmitoyltransferase II (CPTII), carnitine acetyltransferase (CrAT) and choline acetyltransferase (ChAT). Positions related to enzymatic activity are indicated: catalytic His (a), Asp/Glu (b) and the Ser-Thr-Ser motif (c). Tree-determinants conferring differential binding properties are also highlighted: TET/VDN motif (d) in the case of carnitine *vs*. choline preferences, Met *vs*. Ser (e) responsible for sensitivity towards malonyl-CoA and Gly *vs*. Met (f) differentiating long and short acyl-CoAs. B. phylogenetic relationships among acyltransferases and their role in selectivity towards malonyl-CoA. C. Effect of M564G mutation in the CrAT structure. Gly accommodates longer chains, changing the substrate specificity of the enzyme.

elastases (Gly). Two positions adjacent to this in space were originally described as defining substrate differences in these three families: Positions 216 and 226 (in trypsin) are generally glycine in chymotrypsins and trypsins, but replaced by valine and threonine in elastases. Within the trypsin family, the binding site can be divided in several subsites (Fig. (**4**)): The deep hydrophobic S1 pocket, where the conserved Asp189 forms a salt bridge with positively charged moieties. The catalytic triad, or S2 pocket, formed by residues His57, Asp102, and Ser195. The S3 binding subsite, consisting of Gly216. The thrombin insertion loop, Tyr60A-Pro60B-Pro60C-Trp60D (positions 83-86), which forms the hydrophobic P (proximal) pocket. And the hydrophobic distal S4 region (also called D pocket), lined by residues 99, 174, Trp215, and Gly217 (positions 132, 215, 263, and 265 respectively).

Hannenhalli & Russell [11] applied their functional sub-type prediction method to a multiple sequence alignment of elastase, chymotrypsin and trypsin serin-protease sub-types. Given the alignment and some definition of function, such as enzymatic specificity, this method tries to identify positions indicative of functional differences by comparison of sub-type specific sequence profiles. The top two scoring positions identified by the method (position 189, in bovine trypsin, PDB code 5ptp, Z=5.6 and 226, Z=3.9) correspond to two of the pocket positions. The third pocket position (216) has a low Z score (1.0). Inspection of the alignment shows that glycine is frequently tolerated in the elastases sub-type, giving a low Z-score. The third best scoring position (221, Z=3.6) is an Asn residue in elastases, and generally an Ala residue in trypsins, located near to the specificity pocket discussed above. Of the other three positions identified only position 184 (Z=3.1) is near to the other pocket. Here glycine, which is preferred in the elastases may aid the recognition of small side-chains in elastase substrates.

On the other hand, Kastenholz *et al*. [28] studied the problem of ligand selectivity in serin-proteases using the GRID/CPCA method. The results for the GRID/CPCA selectivity were found to be in excellent agreement with the experimental data on selectivity in the thrombin/ trypsin/factor Xa system. Thus, the method finds that the S1 pocket, in spite of having mainly conserved residues, can be used on its own to drive selectivity. In the P pocket, the method predicts that hydrophobic moieties should enhance selectivity for thrombin. The D pocket can also be used to drive selectivity of potential protease inhibitors, especially for the design of selective factor Xa inhibitors, which can take advantage of the interaction properties of the hydrophobic box in factor Xa to accommodate cationic functional groups. Additional residues in the primed side were selected, but not discussed.

Murcia *et al*. have applied COMBINE analysis [37,67] to explain the origin of selectivity of a series of amidinophenylalanines binding to Thrombin, fXa and trypsin (Murcia *et al*. submitted) (Fig. (**5**)). COMBINE analysis selects residues in the S1 pocket, as well as the D-pocket; the S2 region; the interaction with the backbone through residue Gly216 (S3 pocket), and the thrombin exclusive 60-loop (P pocket). COMBINE suggests that selectivity can be largely explained by a handful of VDW interactions, along with a few desolvation energies. For example, the model predicts that ligands with hydrophobic moieties occupying the P pocket should be selective towards thrombin. Similarly, negative charge density in the neighborhood of position 88 (a lysine insertion in thrombin) should be a determinant for thrombin recognition. On the

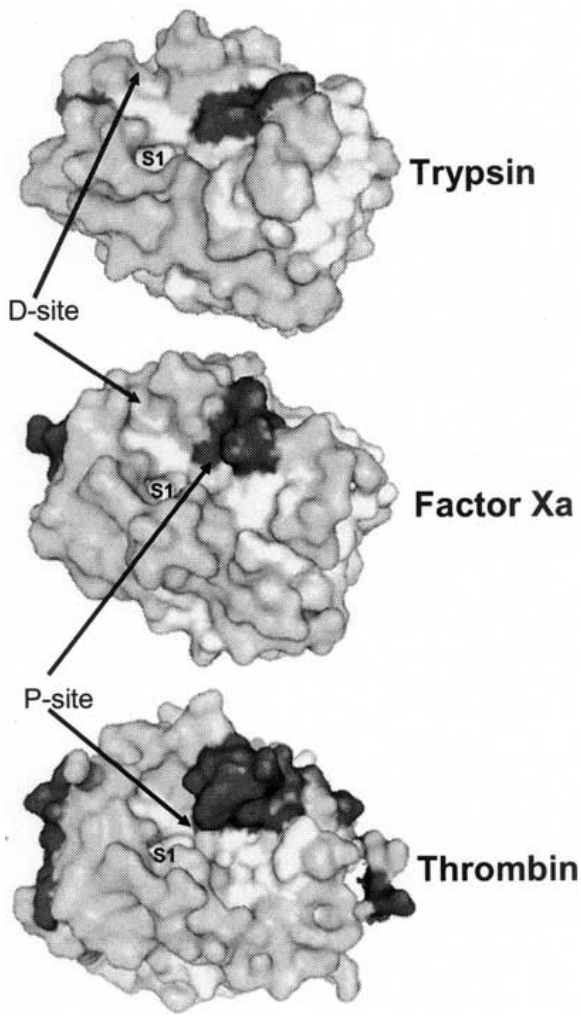

Fig. (4). An overview of the binding site properties of the trypsin-like serine-protease family. The solvent accessible surface for three members -trypsin, factor Xa and thrombin- is shown, colored by the secondary structure of the corresponding residue. Three main subsites of the ligand binding site are highlighted: the S1 subsite, formed by a deep, narrow pocket where the conserved Asp189 forms a salt bridge with positively charged moieties; the D (distal) pocket, lined mainly by aromatic residues and particularly evident in factor Xa, to a minor extent for trypsin, and absent in thrombin; and the P (proximal) pocket, particularly evident in thrombin due to the insertion loop Tyr60A-Pro60B-Pro60C-Trp60D, to a minor extend in factor Xa and absent in trypsin. Two additional subsites, S2 and S3, are not shown for clarity reasons. These are forming the rim of the S1 pocket. See text for additional details. Figure generated with PyMol [87].

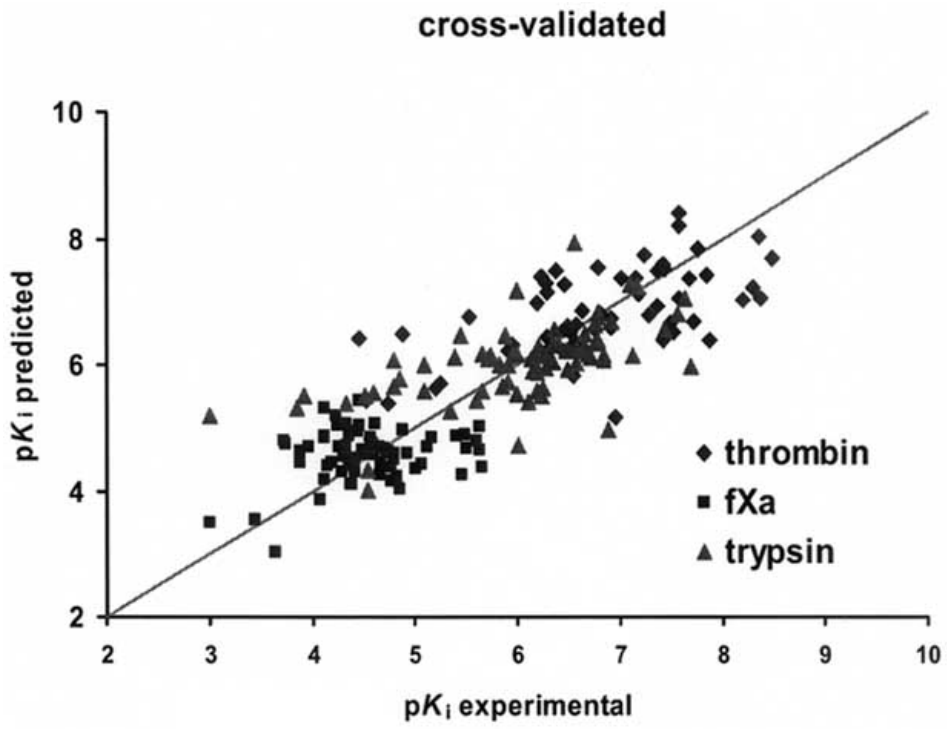

Fig. (5). COMBINE analysis of serin-proteases. Predicted versus experimental affinities for a set of amidinophenylalanines binding to thrombin, factor Xa and trypsin (Murcia *et al*., submitted). See text for details.

other hand, fXa selectivity is enhanced by ligands able to desolvate residues at the entrance of the so-called D-box, and by molecules able to harbor negative charge density in the neighborhood of position 238, where thrombin has a glutamic acid residue, while fXa has a glutamine.

TECHNICAL CONSIDERATIONS

Sequence Analysis

A major problem of sequence space based methods is the quality of the multiple sequence alignments [68,69]. On the one hand, artifacts can be introduced by the heterogeneity of the sequence distributions within the subclasses, with over-represented groups can dominating the alignment [70]. In other cases, protein subfamilies can be composed of sequences too similar or too dissimilar. These deviations from homogeneity can cause practical problems in differentiating residues that are conserved for functional reasons from others that are only apparently conserved due to different artifacts of the distribution of sequences within protein families. An additional problem comes from the quality of the alignment itself. Introduction of structural information in the alignment process has been shown to be beneficial [71].

Modeling Protein Structures

Clearly, both the detection of selectivity sites within protein families and the computational search for putative selective ligands by virtual screening require, in most cases, of the ability to generate high quality homology models for some of the members in the protein family. Recent studies have shown that while homology models are useful in virtual screening, improvements in their quality are still required [72]. CASP competitions have consistently shown that alignment errors and inefficient structural refinement are two of the main factors precluding the derivation of the high accuracy models needed for drug design [73]. Considerable research efforts are underway to resolve these difficulties. Strategies based on the combination of several templates and alignment methods can improve the alignment quality, particularly for remote homologies [74]. Profile-profile alignments, with [22] or without [75] the use of structural information have also been shown to be powerful approaches. On the other hand, structural refinement has been shown to benefit from the use of evolutionary information: applying PCA to multiple structural alignments allows defining a small number of favored evolutionary directions of structural adaptation, where an efficient sampling of the conformational space accessible to the model is possible [76]. Applications of these new structure prediction techniques to the virtual screening scenario have not been reported yet.

When known binders are available, as it happens during the lead optimization step, it is possible to improve the quality of the homology models by demanding self-consistency in the ligand-receptor complexes. Klebe and coworkers have pioneered this approach by refining iteratively homology models using ligand information [77-79]. The procedure starts with the known structure of one or more templates, from which several preliminary homology models of the target are generated. Ligands are then docked into an averaged binding-site representation of the binding-

site models, and new homology models are obtained considering explicitly the docked ligands by transforming the ligand information into user-defined restraints. Ligand-supported homology models are selected as the ones that best explain the observed ligand-binding affinities.

Protein Flexibility in Docking

Improvements in our ability to model selectivity are necessarily is based on our ability to faithfully model the protein-ligand recognition process. Protein flexibility is essential in this process. Initial attempts to consider protein flexibility in docking used modified energy functions with soft van der Waals interactions (soft docking). However, it has been shown that this approximation is too crude, and that a more realistic treatment is to consider explicitly the conformational degrees of freedom of the receptor [80]. During the last few years a deeper understanding of how these conformational degrees of freedom change during the "*induced-fit"* process has started to emerge. Zoete *et al*. [81], for example, compared, for the HIV-1 protease, fluctuations of selected regions calculated with molecular dynamics simulations [82] and normal mode analysis [83], and compared them with knowledge-based fluctuations computed from a set of experimental X-ray structures, focusing mainly on the backbone. The root-mean-square differences observed for the set of structures were shown to have the same variation with residue number as those obtained from molecular dynamics simulations and normal mode analyses, suggesting that both theoretical methods can be useful to model these shifts. In a complementary study, Zavodszky *et al*. [84] selected complexes that do not undergo major main-chain conformational changes upon ligand binding, and studied instead the side chain fluctuations. The authors found that most side chains do not shift to a new rotamer, and that only small motions are both necessary and sufficient to predict the correct binding orientation of the ligand for most complexes in their dataset. These and similar studies are providing insights for the development of better models to account for receptor flexibility in docking, most of them using discrete receptor conformations. For example, Cavasotto & Abagyan [85] have proposed to use a discrete set conformations that consider both side-chain rearrangements and essential backbone movements, and performed flexible ligand-rigid receptor docking and scoring followed by a merging and shrinking step to condense the resulting multiple virtual screening lists in a single one. Similarly, Wei *et al*. [86] considered a method that treats multiple flexible regions of the binding site independently, recombining them to generate different discrete conformations to be used in docking. Their approach improved enrichment of known ligands when a receptor conformational energy weighting term was included in the scoring function, pointing to the need to consider the protein energetics if the receptor degrees of freedom are allowed to vary. This can be implicitly considered by structural sampling with molecular dynamics in absence of the ligand. Meagher & Carlson [87] used this approximation to build pharmacophores for the free HIV-1 protease. The pharmacophore models successfully discriminated known inhibitors from drug-like non-inhibitors.

FUTURE DIRECTIONS

Systems Biology

Pathway information could become useful in the future to help predicting drug-specificity problems [88]. A good example of its potential is the case of the phosphodiesterase (PDE) inhibitor Viagra (Sidenafil). Originally designed to target PDE-5 and promote relaxation of smooth muscle, the compound also binds to the homologous PDE-6 in the eye, which leads to *"blue vision"* in patients, a well-documented side effect difficult to detect in animals. However, sequence searches in pathway collections are able to find these two enzymes as cross-related and are able to infer the likely effects of blocking them [88]. If situations like this one are identified early, efforts can be made to design more selective compounds, and thus potentially avoid problems because of this cross-reactivity. Adequately combined with the computational tools discussed in this review, these approaches could in the future be used not only to find putative competitors, but also to predict the likely biological result of such cross-reactivity.

Chemogenomics

One of the major bottlenecks for the consideration of selectivity early on in a drug discovery research program is the requirement to measure ligand affinities in a high-throughput manner, for a large number of ligands and in a large number of targets. In an important step forward, Fabian *et al*. [89] have recently described an efficient way to experimentally generate systematic small molecule-protein interaction maps across a large number of related proteins. The key of this new technique is to make use of phage-tagged proteins to circumvent the need for conventional protein production and purification. The tagged proteins and immobilized *"bait"* ligands are combined with a *"free"* test compound. If the test compound does not compete efficiently with the *"bait"* ligand for the protein, the tagged protein remains bound to the solid support. The amount of fusion protein bound to the support is measured by quantitative PCR using the phage DNA as a template. As an application, the authors profiled 20 kinase inhibitors against a panel of 119 protein kinases. Interestingly, the authors found that specificity varies widely and is not strongly correlated neither with chemical structure of the ligand nor the identity of the intended target: specificity was shown to vary substantially even among compounds based on the same chemical scaffold, and at the same time, there were many examples of off-targets not closely related by sequence and function to the intended one. This ability to rapidly screen compounds against multiple targets in parallel, blended with the computational methods for specificity profiling discussed in this review, should greatly facilitate and accelerate the drug development process.

CONCLUDING REMARKS

We have surveyed a variety of computational approaches to study the problem of ligand selectivity from different angles. A common theme is the study of factors related to the differential interactions of the ligands with their binding sites, and how these can be unveiled and modeled. However, it is obvious that the target selectivity of a ligand is not only

controlled by details of these interactions, but also involves other factors such as pharmacokinetics, protein environment, differential gene expression, etc… Since, it is helpful to characterize the peculiarities of drug-receptor interactions in the family or subfamily of the target, and use these insights together with other approaches to model more selective ligands. Our survey suggests that the structure-based approach is maturing rapidly and will likely play a significant role in the near future.

ACKNOWLEDGMENTS

ALM is a FPI predoctoral fellow. ELV is recipient of a fellowship from *Fundación Ramón Areces*. MM is grateful to *Fundación Ramón Areces* for a postdoctoral fellowship. AM is the recipient of a postdoctoral contract from *Comunidad de Madrid*. Research at CBMSO is supported by grants from MCYT and FIS, and by an institutional grant from *Fundación Ramón Areces*.

REFERENCES

[1] Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365-370.

[2] Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912-1934.

[3] Cohen, P. Protein kinases--the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **2002**, *1*, 309-315.

[4] Bain, J.; McLauchlan, H.; Elliott, M.; Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem. J.* **2003**, *371*, 199-204.

[5] Matter, H.; Schudok, M. Recent advances in the design of matrix metalloprotease inhibitors. *Curr. Opin. Drug Discov. Devel.* **2004**, *7*, 513-535.

[6] Walker, B.; Lynas, J. F. Strategies for the inhibition of serine proteases. *Cell. Mol. Life Sci.* **2001**, *58*, 596-624.

[7] Coghlan, M. J.; Elmore, S. W.; Kym, P. R.; Kort, M. E. The pursuit of differentiated ligands for the glucocorticoid receptor. *Curr. Top Med. Chem.* **2003**, *3*, 1617-1635.

[8] Takeda-Shitaka, M.; Takaya, D.; Chiba, C.; Tanaka, H.; Umeyama, H. Protein structure prediction in structure based drug design. *Curr. Med. Chem.* **2004**, *11*, 551-558.

[9] Casari, G., Sander, C.; Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **1995**, *2*, 171-178.

[10] Lichtarge, O.; Bourne, H. R.; Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **1996**, *257*, 342-358.

[11] Hannenhalli, S. S.; Russell, R. B. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **2000**, *303*, 61-76.

[12] Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755-763.

[13] Mirny, L. A.; Gelfand, M. S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **2002**, *321*, 7-20.

[14] Mirny, L. A.; Gelfand, M. S. Using orthologous and paralogous proteins to identify specificity determining residues. *Genome Biol.* **2002**, *3*, PREPRINT0002.

[15] Johnson, R. W. D. *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle City, New Jersey, **1998**.

[16] Bauer, B.; Mirey, G.; Vetter, I. R.; Garcia-Ranea, J. A.; Valencia, A. *et al.* Effector recognition by the small GTP-binding proteins Ras and Ral. *J. Biol. Chem.* **1999**, *274*, 17763-17770.

[17] Armon, A.; Graur, D.; Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **2001**, *307*, 447-463.

[18] Landgraf, R.; Fischer, D.; Eisenberg, D. Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* **1999**, *12*, 943-951.

[19] Landgraf, R.; Xenarios, I.; Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **2001**, *307*, 1487-1502.

[20] Yao, H.; Kristensen, D. M.; Mihalek, I.; Sowa, M. E.; Shaw, C. *et al.* An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **2003**, *326*, 255-261.

[21] Pupko, T.; Bell, R. E.; Mayrose, I.; Glaser, F.; Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **2002**, *18 Suppl 1*, S71-77.

[22] Al-Lazikani, B.; Sheinerman, F. B.; Honig, B. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14796-14801.

[23] Winn, P. J.; Religa, T. L.; Battey, J. N.; Banerjee, A.; Wade, R. C. Determinants of functionality in the ubiquitin conjugating enzyme family. *Structure (Camb)* **2004**, *12*, 1563-1574.

[24] Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144-1149.

[25] Basu, G.; Sivanesan, D.; Kawabata, T.; Go, N. Electrostatic potential of nucleotide-free protein is sufficient for discrimination between adenine and guanine-specific binding sites. *J. Mol. Biol.* **2004**, *342*, 1053-1066.

[26] Cruciani, G.; Goodford, P. J. A search for specificity in DNA-drug interactions. *J. Mol. Graph.* **1994**, *12*, 116-129.

[27] Pastor, M.; Cruciani, G. A novel strategy for improving ligand selectivity in receptor-based drug design. *J. Med. Chem.* **1995**, *38*, 4637-4647.

[28] Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E.; Fox, T. GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* **2000**, *43*, 3033-3044.

[29] Naumann, T.; Matter, H. Structural Classification of Protein Kinases Using 3D Molecular Interaction Field Analysis of Their Ligand Binding Sites: Target Family Landscapes. *J. Med. Chem.* **2002**, *45*, 2366-2378.

[30] Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337-344.

[31] Chuaqui, C.; Deng, Z.; Singh, J. Interaction Profiles of Protein Kinase-Inhibitor Complexes and Their Application to Virtual Screening. *J. Med. Chem.* **2005**, *48*, 121-133.

[32] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. *et al.* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889-897.

[33] Wang, J.; Kang, X.; Kuntz, I. D.; Kollman, P. A. Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. *J. Med. Chem.* **2005**, *48*, 2432-2444.

[34] Rizzo, R. C.; Toba, S.; Kuntz, I. D. A molecular basis for the selectivity of thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized born molecular dynamics simulations. *J. Med. Chem.* **2004**, *47*, 3065-3074.

[35] Laitinen, T.; Kankare, J. A.; Perakyla, M. Free energy simulations and MM-PBSA analyses on the affinity and specificity of steroid binding to antiestradiol antibody. *Proteins* **2004**, *55*, 34-43.

[36] Huo, S.; Massova, I.; Kollman, P. A. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J. Comput. Chem.* **2002**, *23*, 15-27.

[37] Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681-2691.

[38] Cramer, R. D., 3rd; Patterson, D. E.; Bunce, J. D. Recent advances in comparative molecular field analysis (CoMFA). *Prog. Clin. Biol. Res.* **1989**, *291*, 161-165.

[39] Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130-4146.

[40] Bhongade, B. A.; Gouripur, V. V.; Gadad, A. K. 3D-QSAR CoMFA studies on trypsin-like serine protease inhibitors: a comparative selectivity analysis. *Bioorg. Med. Chem.* **2005**, *13*, 2773-2782.

[41] Bohm, M.; St rzebecher, J.; Klebe, G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458-477.

[42] Kontogiorgis, C. A.; Papaioannou, P.; Hadjipavlou-Litina, D. J. Matrix metalloproteinase inhibitors: a review on pharmacophore mapping and (Q)SARs results. *Curr. Med. Chem.* **2005**, *12*, 339-355.

[43] Wolohan, P.; Reichert, D. E. CoMFA and docking study of novel estrogen receptor subtype selective ligands. *J. Comput. Aided Mol. Des.* **2003**, *17*, 313-328.

[44] Baskin, II; Tikhonova, I. G.; Palyulin, V. A.; Zefirov, N. S. Selectivity fields: comparative molecular field analysis (CoMFA) of the glycine/NMDA and AMPA receptors. *J. Med. Chem.* **2003**, *46*, 4063-4069.

[45] Hanks, S. K. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol.* **2003**, *4*, 111.

[46] Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A. *et al.* Kinomics—structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243-257.

[47] Wang, Z.; Canagarajan, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H. *et al.* Structural basis of inhibitor selectivity in map kinases. *Structure* **1998**, *6*, 1117– 1128.

[48] Gum, R. J.; McLaughlin, M. M.; Kumar, S.; Wang, Z.; Bower, M. J. *et al.* Acquisition of sensitivity of stress-activated protein kinases to the p38 inhibitor, sb 203580, by alteration of one or more amino acids within the atp binding pocket. *J. Biol. Chem.* **1998**, *273*, 15605– 15610.

[49] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.

[50] McGarry, J. D.; Brown, N. F. The mitochondrial carnitine palmitoyltransferase system. From concept to molecular analysis. *Eur. J. Biochem.* **1997**, *244*, 1-14.

[51] Esser, V.; Britton, C. H.; Weis, B. C.; Foster, D. W.; McGarry, J. D. Cloning,sequencing, and expression of a cDNA encoding rat liver carnitine palmitoyltransferase I. Direct evidence that a single polypeptide is involved in inhibitor interaction and catalytic function. *J. Biol. Chem.* **1993**, *268*, 5817–5822.

[52] Yamazaki, N.; Shinohara, Y.; Shima, A.; Terada, H. High expression of a novel carnitine palmitoyltransferase I like protein in rat brown adipose tissue and heart: isolation and characterization of its cDNA clone. *FEBS Lett.* **1995**, *363*, 41–45.

[53] Price, N. T.; van der Leij, F. R.; Jackson, V. N.; Corstorphine, C. G.; Thomson, R. *et al.* A novel brain-expressed protein related to carnitine palmitoyltransferase I. *Genomics* **2002**, *80*, 433-442.

[54] Bieber, L. L.; Krahling, J. B.; Clarke, P. R.; Valkner, K. J.; Tolbert, N. E. Carnitine acyltransferases in rat liver peroxisomes. *Arch. Biochem. Biophys.* **1981**, *211*, 599-604.

[55] Bieber, L. L. Carnitine. *Ann Rev Biochem* **1988**, *57*, 261-283.

[56] Zammit, V. A. Carnitine acyltransferases: functional significance of subcellular distribution and membrane topology. *Prog. Lipid Res.* **1999**, *38*, 199-224.

[57] Cronin, C. N. Redesign of Choline Acetyltransferase Specificity by Protein Engineering. *J. Biol. Chem.* **1998**, *273*, 24465-24469.

[58] Morillas, M.; Gomez-Puertas, P.; Bentebibel, A.; Selles, E.; Casals, N. *et al.* Identification of Conserved Amino Acid Residues in Rat Liver Carnitine Palmitoyltransferase I Critical for Malonyl-CoA Inhibition. *J. Biol. Chem.* **2003**, *278*, 9058-9063.

[59] Pazos, F.; Sanchez-Pulido, L.; García-Ranea, J. A.; Andrade, M. A.; Atrian, S. *et al.* (1997). Comparative analysis of different methods for the detection of specificity regions in protein families. In Biocomputing and Emergent Computation, D. Lundh, Olsson, B., Narayanan A., ed. (Singapore, New Jersey, London, Hong Kong, World Scientific), pp. 132-145.

[60] del Sol, A.; Pazos, F.; Valencia, A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **2003**, *326*, 1289-1302.

[61] Herrero, L.; Rubi, B.; Sebastian, D.; Serra, D.; Asins, G. *et al.* Alteration of the malonyl-CoA/carnitine palmitoyltransferase I interaction in the beta-cell impairs glucose-induced insulin secretion. *Diabetes* **2005**, *54*, 462-471.

[62] Morillas, M.; Gomez-Puertas, P.; Roca, R.; Serra, D.; Asins, G. *et al.* Structural model of the catalytic core of carnitine palmitoyltransferase I and carnitine octanoyltransferase (COT): mutation of CPT I histidine 473 and alanine 381 and COT alanine 238 impairs the catalytic activity. *J. Biol. Chem.* **2001**, *276*, 45001-45008.

[63] Morillas, M.; Lopez-Viñas, E.; Valencia, A.; Serra, D.; Gomez-Puertas, P. *et al.* Structural model of carnitine palmitoyltransferase I based on the carnitine acetyltransferase crystal. *Biochem. J.* **2004**, *379*, 777-784.

[64] Morillas, M.; Gomez-Puertas, P.; Rubi, B.; Clotet, J.; Arino, J. *et al.* Structural model of a malonyl-CoA-binding site of carnitine octanoyltransferase and carnitine palmitoyltransferase I: mutational analysis of a malonyl-CoA affinity domain. *J. Biol. Chem.* **2002**, *277*, 11473-11480.

[65] Cordente, A. G.; Lopez-Vinas, E.; Vazquez, M. I.; Swiegers, J. H.; Pretorius, I. S. *et al.* Redesign of carnitine acetyltransferase specificity by protein engineering. *J. Biol. Chem.* **2004**, *279*, 33899-33908.

[66] Krem, M. M.; Rose, T.; Di Cera, E. Sequence determinants of function and evolution in serine proteases. *Trends Cardiovasc. Med.* **2000**, *10*, 171-176.

[67] Perez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.* **1998**, *41*, 836-852.

[68] Notredame, C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* **2002**, *3*, 131-144.

[69] Higgins, D. G.; Taylor, W. R. Multiple sequence alignment. *Methods Mol. Biol.* **2000**, *143*, 1-18.

[70] Heringa, J. Local weighting schemes for protein multiple sequence alignment. *Comput. Chem.* **2002**, *26*, 459-477.

[71] O'Sullivan, O.; Suhre, K.; Abergel, C.; Higgins, D. G.; Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **2004**, *340*, 385-395.

[72] McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895-2907.

[73] Tramontano, A.; Morea, V. Assessment of homology-based predictions in CASP5. *Proteins* **2003**, *53 Suppl 6*, 352-368.

[74] Contreras-Moreira, B.; Fitzjohn, P. W.; Bates, P. A. In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. *J. Mol. Biol.* **2003**, *328*, 593-608.

[75] Marti-Renom, M. A.; Madhusudhan, M. S.; Sali, A. Alignment of protein sequences by their profiles. *Protein Sci.* **2004**, *13*, 1071-1087.

[76] Qian, B.; Ortiz, A. R.; Baker, D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15346-15351.

[77] Evers, A.; Gohlke, H.; Klebe, G. Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *J. Mol. Biol.* **2003**, *334*, 327-345.

[78] Evers, A.; Klebe, G. Ligand-supported homology modeling of g-protein-coupled receptor sites: models sufficient for successful virtual screening. *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 248-251.

[79] Evers, A.; Klebe, G. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J. Med. Chem.* **2004**, *47*, 5381-5392.

[80] Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076-5084.

[81] Zoete, V.; Michielin, O.; Karplus, M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mol. Biol.* **2002**, *315*, 21-52.

[82] Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646-652.

[83] Ma, J. New advances in normal mode analysis of supermolecular complexes and applications to structural refinement. *Curr. Protein Pept. Sci.* **2004**, *5*, 119-123.

[84] Zavodszky, M. I.; Kuhn, L. A. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci.* **2005**, *14*, 1104-1114.

[85] Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209-225.

[86] Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161-1182.

[87] Meagher, K. L.; Carlson, H. A. Incorporating protein flexibility in structure-based drug discovery: using HIV-1 protease as a test case. *J. Am. Chem. Soc.* **2004**, *126*, 13276-13281.

[88] Apic, G.; Ignjatovic, T.; Boyer, S.; Russell, R. B. Illuminating drug discovery with biological pathways. *FEBS Lett.* **2005**, *579*, 1872-1877.

[89] Fabian, M. A.; Biggs, W. H. 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D. *et al*. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329-336.

[90] Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T. J. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, *31*, 3497-3500.

[91] Notredame, C.; Higgins, D. G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205-217.

[92] DeLano. The PyMOL Molecular Graphics System. *DeLano Scientific LLC, San Carlos, CA, USA http://wwwpymolorg*.