Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Molecular Graphics and Modelling 24 (2006) 328–340
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials

Christian Hoppe [a,b], Christoph Steinbeck [b], Gerd Wohlfahrt [a,*]

[a] Orion Pharma, Medicinal Chemistry, P.O. Box 65, FIN-02101 Espoo, Finland
[b] University of Cologne, Cologne University Bioinformatics Center (CUBIC), Zülpicher Str. 47, D-50674 Köln, Germany

Received 4 May 2005; received in revised form 29 August 2005; accepted 29 September 2005
Available online 2 November 2005

Abstract

We describe the application of knowledge-based potentials implemented in the MOE program to compare the ligand-binding sites of several proteins. The binding probabilities for a polar and a hydrophobic probe are calculated on a grid to allow easy comparison of binding sites of superimposed related proteins. The method is fast and simple enough to simultaneously use structural information of multiple proteins of a target family. The method can be used to rapidly cluster proteins into subfamilies according to the similarity of hydrophobic and polar fields of their ligand-binding sites. Regions of the binding site which are common within a protein family can be identified and analysed for the design of family-targeted libraries or those which differ for improvement of ligand selectivity.

The field-based hierarchical clustering is demonstrated for three protein families: the ligand-binding domains of nuclear receptors, the ATP-binding sites of protein kinases and the substrate binding sites of proteases. More detailed comparisons are presented for serine proteases of the chymotrypsin family, for the peroxisome proliferator-activated receptor subfamily of nuclear receptors and for progesterone and androgen receptor. The results are in good accordance with structure-based analysis and highlight important differences of the binding sites, which have been also described in the literature.

Keywords: Molecular fields; Grid-based comparison; Knowledge-based potentials; Nuclear receptors; Protein kinases; Serine proteases

1. Introduction

Selectivity towards a biological target is an important property for a drug candidate in order to minimize potential side effects. Traditionally, this has been achieved by cycles of modification and testing of lead compounds. In the absence of structural information of the protein targets, ligand-based QSAR methods have been used to improve specificity, of which comparative molecular field analysis (CoMFA) [1,2] is one of the most successful. A drawback of this approach is that it requires a set of known active molecules with different specificities and whose three-dimensional structures have to be aligned in a meaningful way.

With the rapidly increasing number of protein structures, knowledge of the three-dimensional arrangement of ligand-binding sites became a valuable tool to guide drug design and to introduce receptor specificity early in the discovery process. Molecular fields derived from protein structures have been used to classify and to compare the binding sites of different related receptors [3–5]. These fields were calculated, e.g. with the GRID program [6] using probes whose interaction energies are defined by empirical force fields. Non-grid-based mapping of protein sites has been performed, e.g. by MCSS [7,8], which optimizes the position and orientation of multiple probes in the binding sites. The computationally more demanding MCSS method gives more details than GRID as additional orientational information is provided [8], but as the probe positions are not fixed here, comparison with related receptors is more complex.

Besides empirical force fields, knowledge-based potentials have been proven to characterise receptor–ligand interactions in an appropriate way [9,10]. The use of empirical packing preferences and knowledge-based potentials to assess preferred binding sites in proteins is a well established concept; some

Abbreviations: AR, androgen receptor; LBD, ligand-binding domain; NR, nuclear receptor; PCA, principal component analysis; PLS, partial least squares; PPAR, peroxisome proliferator-activated receptor; PR, progesterone receptor; RMSD, root mean square deviation

* Corresponding author. Tel.: +358 10 4294786; fax: +358 10 4294682.
E-mail address: gerd.wohlfahrt@orionpharma.com (G. Wohlfahrt).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.09.013

examples of this approach include the work of Thornton and co-workers [11], as well as the IsoStar [12] and SuperStar methods produced by Verdonk and co-workers [13]. The advantage of knowledge-based potentials is their ability to describe complex interactions influenced by entropic effects or many-body interactions, which are difficult to quantify with empirical force fields [14].

After mapping of the binding sites, different methods for comparison can be applied, which usually rely on superposition of related protein structures. Principal component analysis [4] or trend vector methods [15] can be applied to extract relevant differences between the fields of the receptors. The first method identifies the most variable features among all receptors in a reduced descriptor space, while the second one finds contour levels above chance correlations from a vector in the original descriptor space.

Most studies have been focussed on the identification of regions which differ between receptors in order to improve selectivity of a ligand, but regions which are common within a protein family are also of interest for the design of family-targeted libraries or to support identification of privileged substructures.

In the present paper, we describe the application of knowledge-based potentials implemented in MOE [16], which use experimental contact statistics fitted to analytical functional forms to identify specific interactions with a protein structure. These potentials include besides distance-dependent also angle and out-off plane dependent distributions. MOE contact statistics have already successfully been used to help refine results from molecular docking runs [17] on the NSAID/COX-2 system, to aid in biodistribution prediction [18] and to explain inhibitor–protein contacts in insect cytochrome P450 binding sites [19]. We calculated the binding probabilities for a polar and a hydrophobic probe on the intersection points of a grid to allow easy comparison of binding sites of superimposed related proteins. The method is fast and simple enough to simultaneously use structural information of multiple proteins of a target family. Using several structures of the same receptor helps to identify the most important interacting regions, which are, e.g. seen with all ligands. Compared to fields derived from a single protein structure this also reduces spurious results, which could be related to experimental inaccuracy or to flexible side-chains resulting in small differences among crystal structures of the same receptor. The method can be used to rapidly cluster proteins into subfamilies according to the similarity of hydrophobic and polar fields of their ligand-binding sites. Regions of the binding site which are common or differ within a protein family can be identified and analysed. Knowledge about common regions is, e.g. useful for the design of family-targeted libraries and differences can be used to improve selectivity of a ligand.

The field-based clustering method is demonstrated for three protein families containing many pharmaceutically relevant targets: the LBDs of nuclear receptors, the ATP-binding sites of protein kinases and the substrate binding sites of proteases [4,5,20,21].

Three serine proteases from the chymotrypsin family are used in the test set. Two of them, thrombin and factor Xa, are involved in the blood clotting cascade and are therefore important targets in the development of anticoagulant or antithrombotic drugs. Trypsin is a pancreatic enzyme involved in digestion. Therefore, selectivity for thrombin and factor Xa over trypsin would improve bioavailability and minimize side effects [4,22].

Most proteins of the nuclear receptor superfamily (NR) act as ligand-activated transcription factors, but the exact mechanism by which the nuclear receptors affect gene transcription is still poorly understood, as is in many cases the role of the subfamilies and their subtypes [2]. Despite the low sequence identity between the LBDs of different NR subfamilies, all NRs share a similar fold and many can bind a range of similar ligands. Depending if the bound ligand is agonistic or antagonistic, the carboxyl-terminal helix H12 is found in either one or another orientation. In the agonist-bound conformation H12 closes the ligand-binding site and shields it from the solvent, whereas in the antagonist-bound conformation H12 does not close the binding pocket. This leads to rather large differences between the properties of the binding sites in the two conformations. A detailed pairwise comparison is presented for the progesterone and androgen receptor, whose binding sites are very similar. High selectivity for only one of the closely related androgen, progesterone, glucocorticoid or mineralocorticoid receptors is important in order to reduce side effects of drug candidates [23,24].

The superfamily of eukaryotic protein kinases is formed of homologous proteins related by their catalytic domains. Although they may have different regulation modes or substrate specificities, they share a common catalytic core structure, which indicates how phosphate is transferred from the kinase to a hydroxyl group in the protein substrate [5]. Kinases play an important role in diverse biological processes such as controlling, signalling and triggering a broad variety of cellular events. Of pharmaceutical interest is the possibility of inhibiting the ATP binding site [5,25–29]. A problem with this approach is that, besides the different kinase subfamilies, more than 2000 ATP-utilizing proteins are estimated in the human genome.

These examples of NRs, kinases and proteases illustrate that methods for analyzing subfamilies or improving subtype specificity of ligands are important in the development of compounds with fewer side effects.

2. Methods

2.1. Protein structures

The protease dataset consists of 13 protein X-ray structures and was taken from literature [30] (Table 1). Sixty-seven nuclear receptor X-ray structures from three subfamilies [31] were taken from the NucleaRDB [32] (Table 2). The kinase dataset was retrieved from the PDB [33] using the search criteria human, X-ray, resolution equal or lower than 2.5 Å and the datasets from Deng et al. [34] and Naumann and Matter [5] (Table 3). Overall 75 protein kinase structures were chosen.