FEBS Letters 349 (1994) 125–130

FEBS 14262

Amino acid preferences at protein binding sites

Hugo O. Villar*, Lawrence M. Kauvar

Terrapin Technologies, 750-H Gateway Blvd., South San Francisco, CA 94080, USA

Received 11 March 1994; revised version received 13 June 1994

Abstract

An analysis of the amino acid distribution at protein binding sites was carried out using 50 diverse macromolecules for which crystallographic data with a bound ligand are available. The purpose of this study is to determine whether differential trends in amino acid distributions exist at binding sites compared to other regions in the proteins. The results indicate that some residues, particularly Arg, His, Trp and Tyr are substantially more frequent at the binding sites, compared to the number of times these residues are present in proteins generally. These effects go beyond the differences seen comparing surface exposed residues to bulk protein. The resemblance in the residue utilization at the binding sites of unrelated proteins restricts the possible types of interactions with ligands, possibly accounting for the repetition of substructural motifs in chemicals with diverse pharmacological action. Further, the use of these diagnostic features may permit identification of ligand binding pockets in a protein structure deduced from sequence information or from data in the absence of a ligand. Some of these findings complement and extend previously described trends for antibody binding sites.

Key words: Binding site; Amino acid distribution

1. Introduction

Although evolution has used the twenty amino acid building blocks opportunistically in constructing proteins, particular amino acids are over-utilized in specific roles in the structure and function of enzymes and proteins. Certain residues have clear involvement in determining the three-dimensional structure of a macromolecule because they affect the folding topology or the overall protein stability. For instance, cysteines are critical to the structural stability in the case of small proteins [1], while arginines and lysines play a role in increasing enzyme compaction [2]. Prolines are also important to structure, since for the most part, they grossly alter the folding topology or more subtly induce kinks in α-helical domains such as in transmembrane regions [3]. The significant wealth of knowledge regarding the influence of a given amino acid on protein structure has been partly derived from the statistical analysis of the steadily expanding protein crystallographic data [4–6]. Most of the information derived by statistical analysis has been applied to detect sub-cellular distribution signaling sequences, to elucidate unknown structures by homology modeling, to postulate potential protein folding intermediates, and most ambitiously to predict de novo structure from sequence alone.

By contrast, the role of different residues in ligand accommodation and binding has attracted less attention. A study of the amino acid distributions at the complementarity determining regions (CDRs) of available antibody structures revealed that there is a high frequency of participation of Tyr and Trp in these regions involved in antibody–antigen complex formation. The significance of the results becomes particularly clear when they are compared to the average for proteins or even for the averages found for other regions of the antibodies [7]. The differential usage of these residues was proposed to account for the cross-reactivity profiles observed in antibodies. That is, the over-abundance implies that these residues are very important for antigen recognition, but the fact that different antibodies must use the same residues for recognition limits the number of binding motifs, which results in a limit to the specificity that can be achieved.

An equivalent situation to the cross-reactive patterns of antibodies is commonly observed in enzymes, receptors and other proteins that are targets for drug development. Indeed, the ability of antagonists and inhibitors to bind to unrelated macromolecules is well documented, in clear parallel with the cross-reactivity observed for antibodies [8]. If the ability of ligands to bind promiscuously is attributed to implicit similarities in the characteristics of the binding sites, then following the logic applied to antibodies, over-representation of certain amino acids in the recognition sites of proteins should be found not only in the CDRs of antibodies but also at the binding sites of other proteins. Enough structures from diverse classes are now available to test this idea.

In this article, the amino acid distribution surrounding ligands in complexes determined crystallographically are studied to determine if differential trends in amino acid distributions are found. The results provide a general view of the biophysical properties of binding sites, and the utility of certain amino acids for interaction com-

*Corresponding author. Fax: (1) (415) 244-9388.

SSDI 0014-5793(94)00648-F

pared to other possible roles, such as building the protein architecture. Analysis of such *general characteristics* of residues involved in ligand recognition by proteins may shed light on basic aspects common to a wide array of biological interactions.

2. Materials and methods

The amino acid composition was determined from the crystal structures available at the Brookhaven Protein Databank. The values presented were calculated as a function of the distance from a bound ligand for 50 proteins. The structures were selected to avoid a bias for any one group of proteins, or ligands with similar function. For example, while there are data for a large number of serine proteases with bound inhibitors, only three were adopted for the present study. The complete list of enzymes and their bound ligands is given in Table 1.

For each enzyme, the number of each of the twenty amino acids was counted at ten different distances from the ligand. A residue was counted if any heavy atom of the residue was separated from any atom of the ligand by a distance shorter than or equal to the value considered. The results were expressed as percent composition for that distance.

3. Results and discussion

Fig. 1 shows in gray for each amino acid the percent composition separated from the ligand by up to 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 Å respectively, and in black the average found for that amino acid in all fifty proteins studied. A residue was counted if any heavy atom of the residue was separated from any atom of the ligand by a distance shorter than or equal to the value considered. In the same plot, using the right hand scale, the ratio of the frequency at 4 Å relative to the average is shown, as a percentage. This relative frequency represents the propensity of an amino acid to be close to the bound ligand.

Some clear trends in amino acid distribution are revealed by examining the relative frequency of a particular residue at a given distance to the average observed in all proteins. Trp and His are found 250% more frequently in contact with the ligand than their average observed across all proteins. The relative numbers of Arg

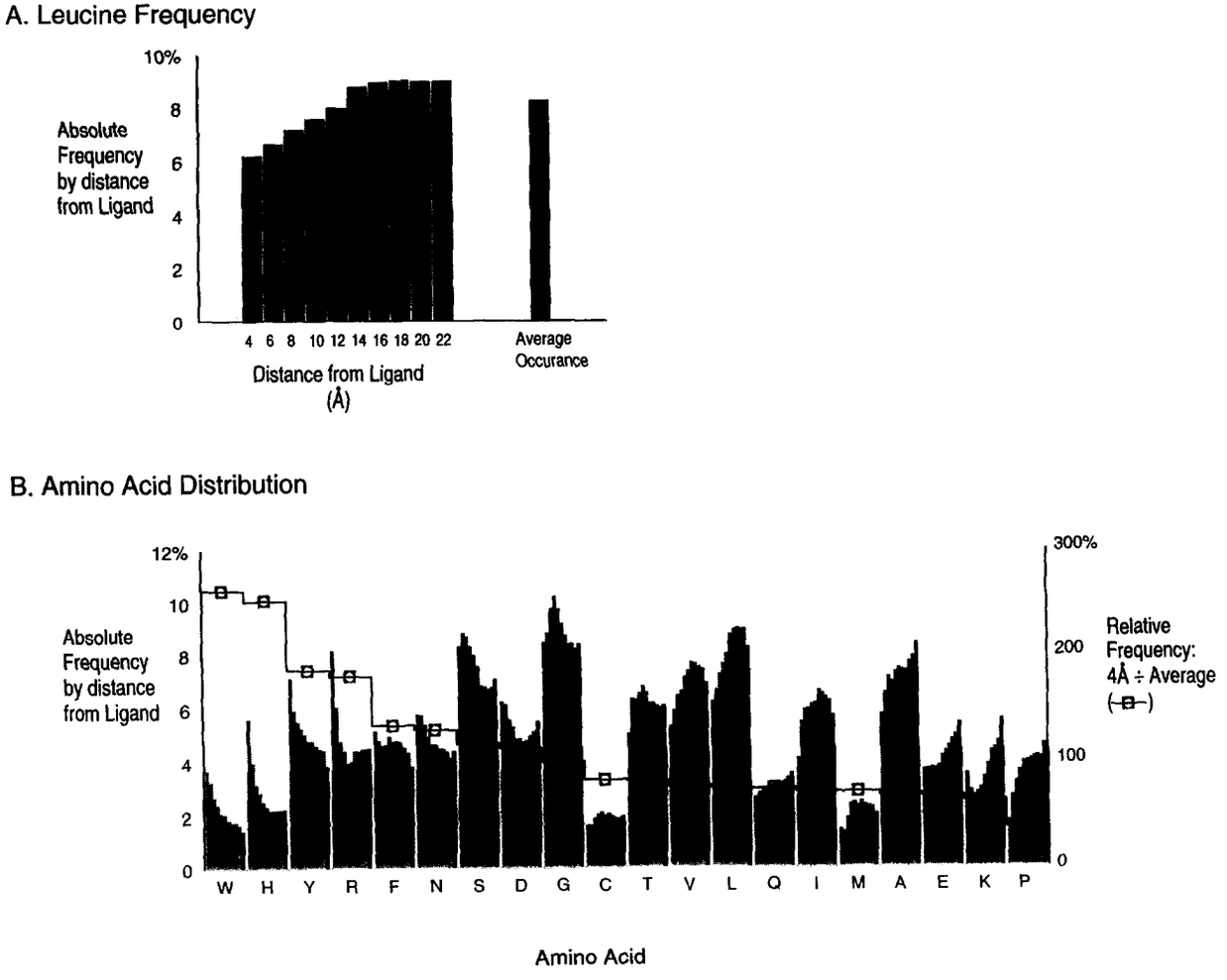

Fig. 1. Amino acid frequency as a function of the distance from ligand binding site. The histograms are built from the percent frequency of the amino acid at 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 Å from the position of the ligand. (a) Exemplifies the diagram for Leu. (b) The same diagram is presented for each amino acid. Each bar corresponds to a 2 Å increment in the distance from the ligand. The overall average values are also shown as a black bar at the right for each amino acid. The ratio of the composition at 4 Å versus the overall average is plotted as a per cent for each amino acid as a square, using the right axis. The amino acids are ordered depending on their over-representation at the binding site.

and Tyr residues also show a significant increase at the active site, close to 200% of the average. More modest increases are found for Ser and Asp. The most significant decrease in proportion at the binding region compared to its overall average occurs for Pro. Other residues that decrease their frequency are Lys, Glu and Ala.

A second way to analyze the data is by observing the *absolute composition* of the active site. Gly, Ser, Arg and Tyr are the residues most abundant at 4 Å from the ligand. These residues are thus also over-represented at the binding sites since the ones most commonly found on the overall average are significantly less polar. Combin-

Table 1
List of the crystallographic structures examined for protein–ligand interactions

Protein	Ligand	Code	EC number
βAlcohol dehydrogenase	NAD	3hud	1.1.1.1
M4 lactate dehydrogenase	Oxamate	1ldm	1.1.1.27
Glutathione reductase	Retro GSSG	4grl	1.4.6.2
Methylamine dehydrogenase	Amicyanin	1mda	1.4.99.3
Dehydrofolate reductase	Folate	1dhf	1.5.1.3
p-Hydroxybenzoate hydrolase	*p*-hydroxybenzoate	2phh	1.14.13.2
Cytochrome P_{450} cam	Camphane	6cpp	1.14.15.1
Ferrodoxin	2′-Phospho-5′-AMP	2fnr	1.18.1.2
Aspartate carbamoyl transferase	Phosphonoacetamide	1atl	2.1.3.2
Chloramphenicol acyltransferase	Chloramphenicol	1cla	2.3.1.28
Glycogen phosphorylase β	Pyridoxal-5′-phosphate	1gpb	2.4.1.1
Glutathione *S*-transferase 3:3	γ-Glu-Cys-Gly	1gst	2.5.1.18
Glutathione *S*-transferase P1-1	γ-Glu-Cys-Gly	1gss	2.5.1.18
Aspartate aminotransferase	α-methylaspartate	1ama	2.6.1.1
Phosphofructokinase	2-Fructose-1,6-biphosphate	1pfk	2.7.1.11
Adenylate kinase-3	AMP	1ak3	2.7.4.10
Acetylcholine receptor	Acetylcholine	1ace	3.1.1.7
Fructobiphosphatase	Fructose 6-phosphate	1fbp	3.1.3.11
Ribonuclease	3′-guanylic acid	1rms	3.1.4.23
Lys-25 ribonuclease	Guanylyl-2′,5′-guanosine	2rnt	3.1.27.3
N9-neuraminidase	FAB-NC41	1nca	3.2.1.18
Barnase	Deoxynucleotide inhibitor	1rnb	3.4.2.15
Carboxypeptidase *a*	Gly-Tyr	3cpa	3.4.17.1
γ-Chymotrypsin	*trans-o*-hydroxy-α-methylcinnamate	3gch	3.4.21.1
Thermolysin	Phosphoramidon	1tlp	3.4.21.4
β-Trypsin	*p*-Amidinophenylpyruvate	1tpp	3.4.21.4
Thermitase	Eglin C	1tec	3.4.21.14
Endothiapepsin	H77	1er8	3.4.23.6
Renin	C60	1rne	3.4.23.15
Rubisco	3-Phosphoglycerate	1rus	4.1.1.39
Citrate synthase	L-Malate	1csc	4.1.3.7
Carbonic anhydrase II	Thiocyanate	2ca2	4.2.1.1
Aconitase	Transaconitase	1aco	4.3.1.3
Triosephosphate isomerase	*N*-Hydroxyphosphonobutanamide	1tsi	5.3.1.1
Xylose isomerase	Xylitol	2xis	5.3.1.5
Bioperon repressor	Biotin	1bib	6.3.4.15
Non-enzymes			
Choleratoxin	Lactose	1ltt	
Concanavallin	α-Methyl mannoside	4cna	
FK-506 binding protein	FK-506	1fkf	
IG MC-PC603	Phosphocholine	2mcp	
Lectin	Lactose	1lte	
Major urinary protein	2sec Butyl thiazoline	1dog	
Myoglobin	Azide	5mba	
Plasminogen kringle4 human	ε-Aminocaproic acid	2pk4	
Ras protein	GDP	1q21	
Rhinovirus coat protein	Win VIII	1r08	
Ricin	Adenylguanosine	1apg	
Streptavidin	Biotin	1stp	
Transthyretin	Milrinone	1lm	
Wheat germ agglutin	2-*N*-Acetylneivaminyllactose	1wgc	

The Code refers to the Brookhaven database code. The enzymes are listed by Enzyme Commission number.

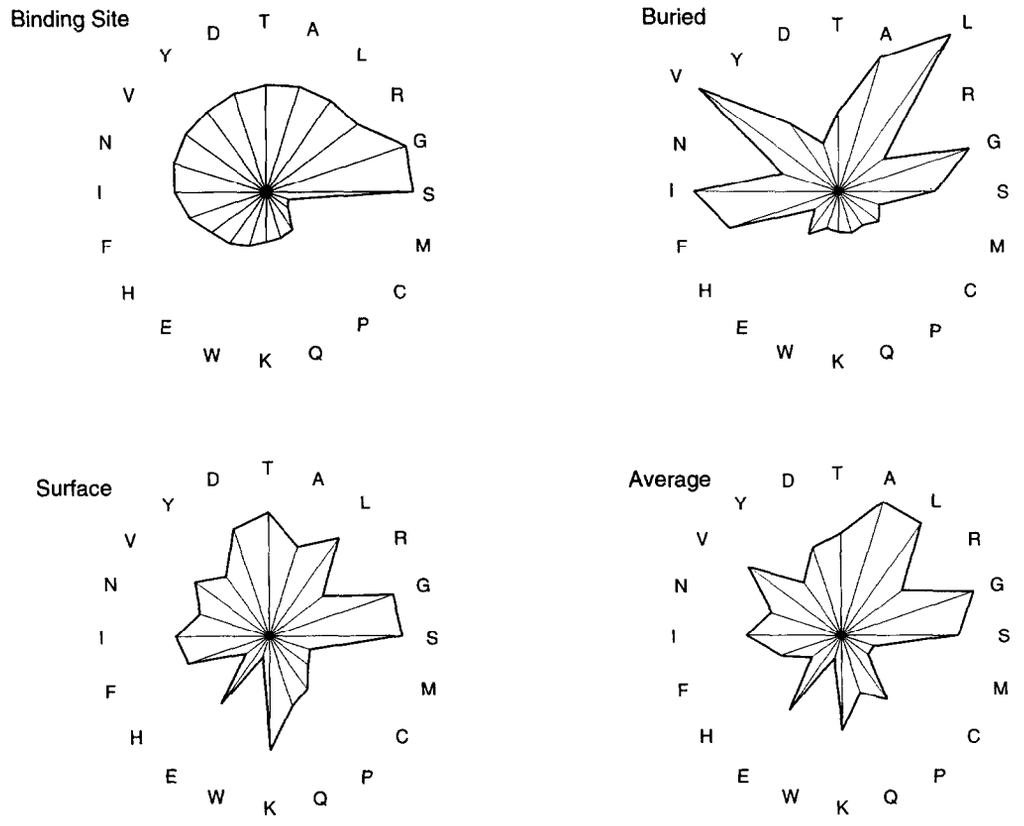

Fig. 2. Star diagram showing the absolute percent amino acid frequency observed at 6 Å from: (a) any point on the ligand; (b) any point on a random residue located on the protein surface; (c) any random residue located in the interior of the protein; (d) average over all the proteins studied. The length of each spoke is proportional to the frequency of that amino acid at that site.

ing the relative and absolute frequency data, the amino acid distribution at protein binding sites has some similarities with the observations made for the antibody CDRs. As in the antibody CDRs, Tyr and Trp are among the residues that show the most significant increase compared to the average found when the entire proteins are considered. In antibody CDRs, however, Trp and Tyr [7] are also the most abundant on an absolute basis, which is not the case for the proteins studied here.

The dominance of Gly, Ser, Arg and Tyr in protein binding sites is likely to have an impact on the secondary and tertiary structures of the elements that constitute the active site. Gly, Ser and Tyr have a tendency to be found in β-turn regions [4], which have been proposed to be significant structural elements defining protein specificity, and to be partly responsible for determining the characteristics of the binding sites [10]. The distributions observed are consistent with such an hypothesis.

Analysis of the structures used shows that in most cases Ser and Tyr are largely involved in hydrogen bond interactions with the ligands. Arg appears to be the residue of choice to form salt bridges, particularly with polyvalent ions. Trp is part of hydrophobic interactions, as in acetylcholine receptor, or π–π stacking complexes as in aspartate aminotransferase.

Since highly polar residues abound on the protein surface, it is not surprising that residues such as Lys and Glu increase their representation at large distances from the binding site. Other polar residues, including Arg, have a minima in their frequency distribution at intermediate distances that is exactly opposite to the peak observed for the apolar residues such as Val, Leu, Met and Ile. The peaking of hydrophobic residues at intermediate distances represents the hydrophobic core present in most proteins.

In a parallel experiment, the results were compared to those found using other starting points instead of the ligands as a reference from which to measure distances. Random residues on either the surface or buried in the protein interior took the place of the ligand in the analysis. The results are shown in Fig. 2. The absolute distributions of amino acids within 6 Å from the reference moiety are significantly different on the surface, buried, or surounding the ligand. Moreover, the overall amino acid average is significantly different from each of the others.

Reassuringly, the results at a short distance from the

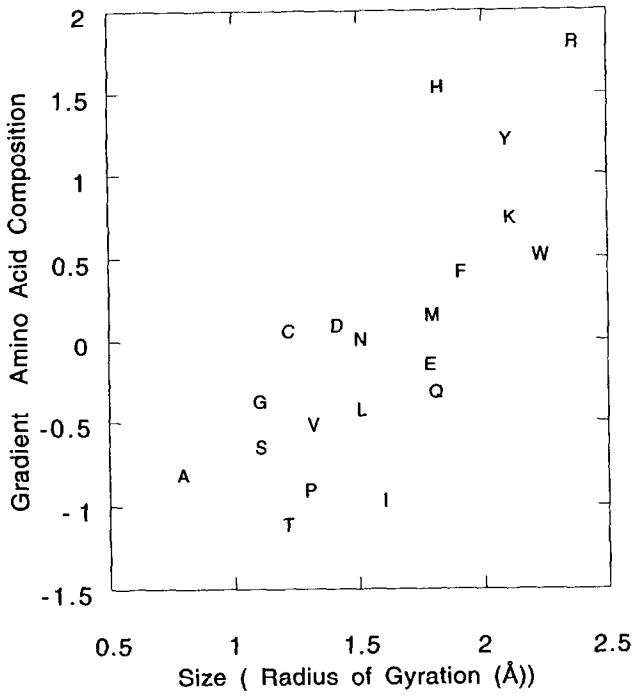

Fig. 3. Gradient in absolute amino acid composition between 4 and 6 Å from the ligand binding site as a function of the radius of gyration for that amino acid.

reference point reflect the hydrophobicity of the location, substantiating the counting algorithm. There is a remarkable abundance of apolar residues near the buried starting points: Leu, Val, Ile and Ala. Trp appears more frequently encircling buried residues, compared to the protein average, but not as frequently as observed at the binding sites. Lys, Pro and Glu appear more frequently on the surface while Trp, His and Arg are more frequent at the binding site. Indeed, the differential distributions noted here provide strong enough discrimination to be tentatively considered as diagnostic of pockets that serve as binding sites when the actual site is unknown.

In addition to electronic effects, steric factors might also be important to the structure of binding sites. To explore this possibility, the *rate of change* in the amino acid composition in the vicinity of the ligand was recorded, using the difference of the compositions observed at 4 and 6 Å as a first approximation to a gradient of utilization and the radius of gyration as a measure of the steric bulk [11]. An inverse linear correlation of the change in composition with the residue radius of gyration was observed (Fig. 2). Larger residues increase in abundance faster than small residues at the binding site. The end result of this trend is that binding sites have more gaps than the densely packed bulk protein, as appropriate for accommodating bound ligands.

Several aspects of the statistical significance of the results were examined. First, the same analyses as before were repeated twenty times on 45 proteins rather than 50, taking out of the set random groups of five enzymes. A small variability of less than 0.5% was observed for the shorter distances from the binding sites, a number significantly smaller than the differences discussed. Longer distances involve larger numbers of residues and show even less sensitivity. Moreover, the results are not particularly sensitive to which proteins are included in the set.

The data were also analyzed looking at distributions confined to concentric shells, e.g. from 4 to 6 Å rather than the cumulative values reported here. Although similar general conclusions could be drawn using the concentric shells, the limited number of structures made statistical arguments weaker without adding any new insights.

4. Conclusions

A protein's structure as revealed by primary sequence and X-ray crystallography reflects a variety of constraints: evolutionary history, folding pathway, stability in native milieu, and function. Considerable progress has been reported in the literature regarding generalizations relating to the first three constraints, but the fourth has largely been viewed as idiosyncratic to each protein. The results reported here suggest that it may be possible to identify functional pockets in a protein's structure by comparison to features of ligand binding sites in general, expanding on the success achieved in predicting antibody structures.

From the perspective of drug design, the resemblance in residue utilization at binding sites for unrelated proteins indicates that limits may exist to the possible types of interactions with other molecules. Consequently, some types of chemical structures should be favored for interaction with a macromolecule. Such limitations may account for the observed presence of particular substructural motifs across pharmacologically diverse classes of chemicals that elicit their action by blocking a recognition site [8]. Alternatively, the appearance of common motifs may only reflect the limited variability of currently available chemical libraries from which drugs are derived. A deeper understanding of the structure of binding sites, and the constraints they impose on the diversity of pharmaceutically useful compounds, will be critical to directing new synthetic chemical efforts, particularly in the case of parallel solid phase synthesis of related molecules [12] for which considerable preparatory effort is required.

References

[1] White, S.W. (1992) J. Mol. Biol. 227, 991–995.
[2] Nandi, C.L. Singh, J. and Thornton, J.M. (1993) Protein Eng. 6, 247–259.
[3] von Heijne, G. (1991) J. Mol. Biol. 218, 499–503.

[4] Chow, P.Y. and Fasman, G.D. (1978) Annu. Rev. Biochem. 47, 251–276.
[5] Benner, S. (1992) Curr. Opin. Struct. Biol. 2, 402–412.
[6] Luthy, R., Mc Lachlan, A.D. and Eisenberg, D. (1991) Proteins 10, 229–239.
[7] Mian, I.S., Bradwell A.R. and Olson, A.J. (1991) J. Mol. Biol. 217, 133–151.
[8] La Bella, F.S. (1991) Biochem. Pharmacol. 42, S1–S8.
[9] Bernstein,F.C., Koetzle,T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[10] Murakami, M. (1993) J. Prot. Chem. 12, 783–789.
[11] Waals, P.H. and Sternberg, M.J.E. (1992) J. Mol. Biol. 228, 277–297.
[12] Gallop, M.A, Barrett, R.W., Dower, W.J., Fodor, S.P.A. and Gordon, E.M. (1994) J. Med. Chem. 37, 1233–1251; Jung, G. and Beck-Sickinger,A.G. (1992) Angew. Chem. Int. Ed. Engl. 31, 367–383.