DOI: 10.1002/minf.201400090

Fractal Dimensions of Macromolecular Structures

Nickolay Todoroff,[a] Jens Kunze,[a] Herman Schreuder,[b] Gerhard Hessler,[b] Karl-Heinz Baringhaus,[b] and Gisbert Schneider*[a]

Abstract: Quantifying the properties of macromolecules is a prerequisite for understanding their roles in biochemical processes. One of the less-explored geometric features of macromolecules is molecular surface irregularity, or 'roughness', which can be measured in terms of fractal dimension (*D*). In this study, we demonstrate that surface roughness correlates with ligand binding potential. We quantified the surface roughnesses of biological macromolecules in a large-scale survey that revealed *D* values between 2.0 and 2.4. The results of our study imply that surface patches involved in molecular interactions, such as ligand-binding pockets and protein-protein interfaces, exhibit greater local fluctuations in their fractal dimensions than 'inert' surface areas. We expect approximately 22% of a protein's surface outside of the crystallographically known ligand binding sites to be ligandable. These findings provide a fresh perspective on macromolecular structure and have considerable implications for drug design as well as chemical and systems biology.

Keywords: Ligand binding · Computational chemistry · Drug discovery · Protein structure · Fractal

1 Introduction

Molecules are commonly thought to interact *via* complementary surfaces, which possess surface properties that allow the formation of stable complexes. Structure-based drug design aims to target such complexes with synthetic compounds to influence the complexes' biological functions in living tissue. To date, the best surface properties for predicting molecular interactions are either physicochemical, such as lipophilicity and charge distribution, or geometric, such as surface concavity and shape.[1] Several attempts have been made to identify interaction "hot spots" and predict the "druggability" or "ligandability" of protein surface regions.[2] It has been realized that pure geometric properties bear relevant information about the nature of a protein-ligand interaction.[3] An intuitive way to understand and quantify the geometric information of a surface using a single number is to estimate its fractal dimension, *D*. A smooth surface is expected to have a lower fractal dimension than a more irregular object.[4] Although this concept was initially suggested three decades ago,[5,6] the lack of large-scale studies on the topic as well as the partially contradictory results have occluded its potential impact on our understanding of macromolecular structure and function. Here, we present a full-fledged analysis of the fractal dimensions of both macromolecular surfaces and low-molecular-weight drugs. The results of our study pinpoint surface roughness as a local, dynamic, and context-dependent molecular property, and they highlight the great relevance of this decisive feature of molecular interaction sites to the discovery of innovative drugs and chemical probes as well as for their modes of action.

2 Results and Discussion

To calculate the perceived roughness of a molecular surface, we created a discrete approximation of the molecules' solvent-excluded surface (*SES*) by sampling approximately equidistant points on the *SES*. We worked with a density of seven vertices per Å^2 and a rolling probe radius of 1.5 Å while ignoring hydrogen atoms for the *SES* calculations.[7] We chose the spatial correlation dimension[8] to calculate the fractal dimension and, thus, the perceived roughness of the molecular surfaces. The correlation integral $C(\delta)$ is a measure of the spatial distribution of surface points X in their embedding space. The integral is estimated by the correlation sum $C(X, \delta)$ (Equation 1), which is used as a measurement at fractal scale δ. In turn, δ is the radius of the hypersphere used to define the neighborhood of a surface point x. D was calculated by computing $C(X, \delta)$ for the δ values 0.4, 0.8, 1.6, and 3.2 Å. In that calculation, the lowest radius used was considerably greater than the shortest dis-

[a] N. Todoroff, J. Kunze, G. Schneider
Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences
Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland
fax: (+41) 44 633 13 79
*e-mail: gisbert@ethz.ch

[b] H. Schreuder, G. Hessler, K.-H. Baringhaus
Sanofi-Aventis Deutschland GmbH R&D
Frankfurt am Main, Germany

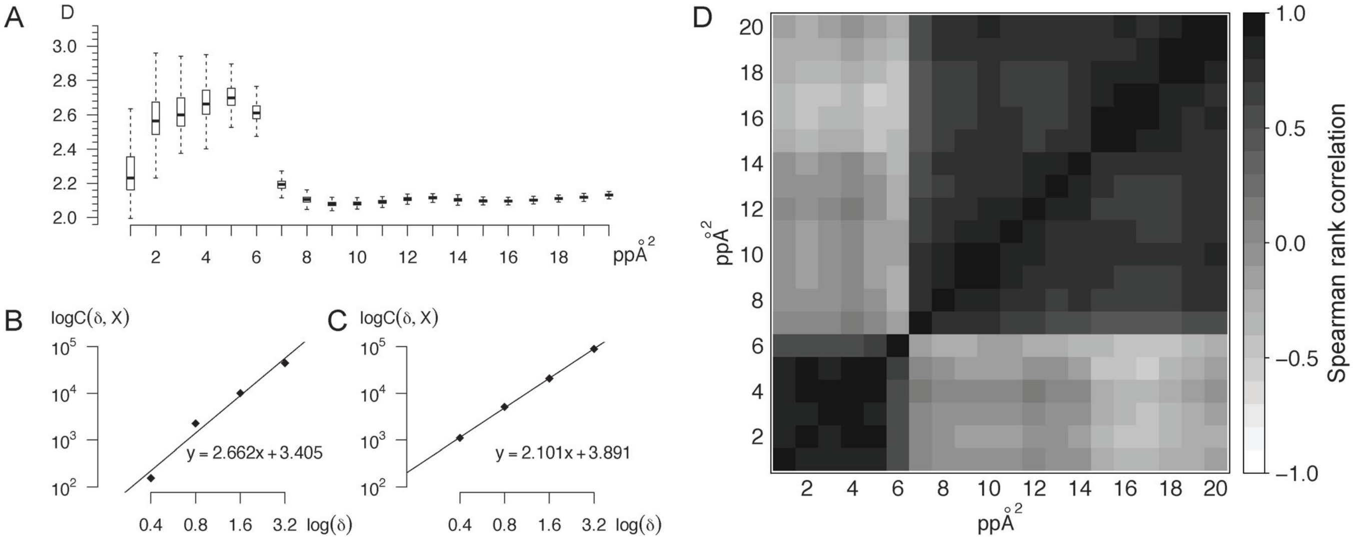

Figure 1. Discrimination ability of the fractal dimension *D* depending on the fractal scale and point density (ppÅ^2: points per Å^2) calculated for the 604 ligands of the main set of protein-ligand complexes. A) Dispersion of *D*. A tendency of decreasing dispersion is apparent with increasing discretization density. A steep fall in *D* values as well as their dispersion manifests itself at density > 6 ppÅ^2. B, C) Regression lines of the doubly logarithmic plot of the correlation sum for the ligand of the PDB ID: 1c5z complex for densities on both sides of the steep fall in *D*: B) 5 ppÅ^2, (r^2 = 0.952); C) 7 ppÅ^2, (r^2 = 0.999). D) Gray scale coded matrix containing the Spearman's rank correlation of the vectors of *D* values of the ligands (one 604-dimensional rank vector for each density). Each filled square represents the pairwise rank correlation between two corresponding vectors, illustrating how the change in density shuffles the relative positioning of the *D* values. Few changes in the ligands' positions in the rank vector indicate stable discrimination ability (strong correlation, dark color).

tance between any two adjacent points on the surface, and the largest radius was of the same magnitude as the average diameter of an amino acid side chain. In Equation 1, Θ is the Heaviside step function, which equals zero if the distance between surface points x_i and x_j exceeds δ.

$$D \approx \lim_{\delta \to 0} \frac{\log C(X, \delta)}{\log \delta}, \qquad (1)$$

where

$$C(X, \delta) \equiv \frac{1}{N^2} \sum_{i,j=1}^{N} \Theta\left(\delta - \left\| x_i - x_j \right\|\right).$$

Using this setup, we conducted a large-scale analysis of crystallographically obtained structures, including DNA (n = 786), RNA (n = 449), proteins (n = 604), and protein-bound drug-like ligands (n = 604), which were acquired from the Protein Data Bank (PDB).[9] All of the molecules exhibited *D* values greater than 2.0, which is the value of an idealized flat surface (D_{plane} = 2.0), but smaller than the value for a solid sphere (D_{sphere} = 3.0) (Figure 2). The molecules all tended toward increasing *D* values as the sizes and degrees of freedom increased. In general, the biological macromolecules were rougher than their ligands. This observation could be explained by their substantial differences in size and flexibility. Among the macromolecules, DNA structures have the fewest degrees of freedom due to their relatively rigid backbones, followed by RNA with its single strands, loops, and bulges. Proteins have the greatest conformational freedom among these three types of macromolecules. The observed *D* values of the corresponding ligands were the most diverse among the inspected molecule classes.

Local surface properties are of particular interest for protein-protein and protein-ligand interactions, and those interactions are of great relevance for pharmaceutical research, chemical biology and chemogenomics.[10] Thus, we also extracted and calculated *D* values for four different types of protein surface patches based on their locations (Figure 2). Randomly sampled surface patches exhibited the lowest roughness and had the highest average spreads. Protein surface cavities, as calculated by PocketPicker,[11] yielded higher average *D* values, while the *D* values of known ligand-binding pockets and protein-protein interfaces were the highest overall. This result suggests that the roughness of a surface patch is either involved in the formation and stabilization of the protein complex or is a consequence of it. Our results support studies implying that the perceived roughness of protein-protein interfaces is higher than average,[5,6] and this paper challenges the conclusion that enzymatic active sites are smoother than the average protein surface.[5] The median values observed for low-molecular-weight ligands (*D* = 2.193) and for known ligand-accommodating pockets (*D* = 2.199) were similar; however, we found that there is no pairwise correlation between the ligands and either their respective pockets (R^2 = 0.006) or their bound proteins (R^2 = 0.002). This observation

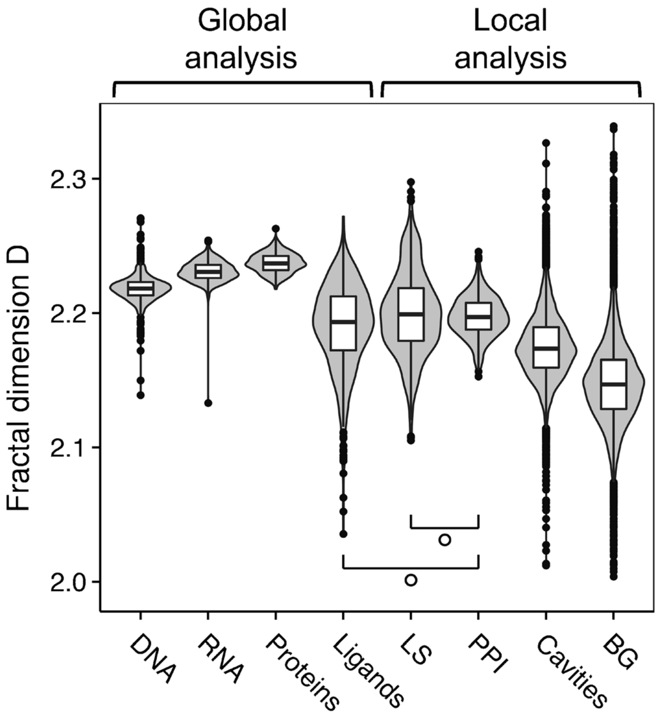

Figure 2. Dispersion analysis violin plots of the fractal dimension D of different biomolecular surfaces, including complete molecules (DNA, RNA, proteins, and ligands) and protein surface patches. Also shown are the liganded protein sites (LS), bound protein–protein interfaces (PPI), cavities on the protein surface (Cavities), and the background, which was computed from randomly sampled protein surface patches (BG). The symbol '○' indicates statistically insignificant differences between the distributions. We used a nonparametric test (Mann-Whitney-U). To compensate for α-errors, the p values of the 28 independent tests were Bonferroni-corrected. The null hypothesis could not be rejected for the Ligands-PPI, and LS-PPI pairs implying similarities in their distributions. The sample sizes were as follows: DNA ($n=786$); RNA (449); Proteins (604); Ligands (604); LS (604); PPI (466); Cavities (12,543); BG (11,953).

contradicts the hypothesis that a pair of binding partners should have similar or opposite perceived roughness values from a statistical point of view, which was an unexpected result.

While pursuing the answer to whether the perceived roughness could pinpoint the locations of preferred ligand-binding sites, we replaced the concept of predefined pockets on the molecular surface with a neighborhood-around-an-atom approach (Figure 3A). We investigated the relevance of *D* for atoms close to known ligand-binding sites as opposed to atoms that were located farther away. To acquire comparable surface patches for each atom, we used a sampling method that tallies the contributions to the *SES* of all of an atom's neighbors within a radius of 6 Å. We then calculated the *D* for each corresponding atom set and gathered statistics about all of the *D* values for the atoms within the distance threshold. Each obtained statistic for a neighborhood was assigned to the atom in the center of the hypersphere defining the neighborhood's limits. We observed a remarkable difference between the distributions of *D* for interacting atoms and the distributions of *D* for atoms farther away from the ligand-binding zone (Figure 3A, right panel). On average, binding sites contained atoms with both high (*rough* neighborhoods) and low *D* values (*smooth* neighborhoods). However, this effect was not observed across the entire pockets dataset. Evidently, a surface's *smoothness,* in addition to its *roughness*, provides relevant information for detecting regions of interest for ligand-receptor interactions when specifically considering potential binding sites outside of the surface cavity context.

Based on these insights, we devised a local roughness (*LR*) model that operated by gathering information from the neighborhoods of surface-exposed atoms. The *D* values of all of the atoms within a predefined radius *r* surrounding a reference atom *a* constituted a set, i.e., $LR_a(r)=\{D_b \mid ||b-a|| < r\}$. An investigation of the distribution of each $LR_a(r)$ set revealed a tendency toward increasing means and decreasing standard deviations for *D* as the distance from *a* to the ligand binding site increased (Figure 3B). Negative control calculations were carried out by randomly selecting atoms on the proteins' surfaces to serve as mock ligand-binding sites. We observed differences between atom sets of the actual interaction sites and the sets of the atoms that were randomly chosen, indicating that the neighborhood distribution of *D* contained a non-negligible discriminative power, i.e., a random group of atoms did not mimic the behavior of atoms surrounding known ligand-binding sites. The notable differences found in the proteins' roughness features were supported by a negative pairwise correlation between the binding-site data and the random controls ($\sigma(LR_a(r))$: $R=-0.43$, $R^2=0.19$; $\overline{LR_a(r)}$: $R=-0.65$, $R^2=0.42$). To test our hypothesis, we conducted a positive control calculation in which we set the individual atoms closest to the geometric centers of the corresponding ligands as the centers of the neighborhood spheres. This test did not fail. Instead, the strong correlation observed between the ligand atom-derived distribution of the $LR_a(r)$ values and this coarser ligand center-derived distribution corroborated our conclusions $\sigma(LR_a(r))$: $R=0.97$, $R^2=0.95$; $\overline{LR_a(r)}$, $R=0.90$, $R^2=0.82$).

Finally, we incorporated this information into a computerized learning experiment that was used to classify protein atoms participating in ligand binding sites (potential binding "hot spots"). All of the protein atoms within 3 Å of a ligand atom were considered atoms of interest (positive samples) and were compared with the background containing all of the other protein atoms (negative samples) (Table 1). In this experiment, the classifier function calculated the probability that a surface patch was ligandable. A patch was considered ligandable if it received a predicted probability greater than 50 %.

The best performance, in terms of separating the atoms of interest from other atoms while balancing accuracy and early retrieval, was achieved using the nearest-neighbor

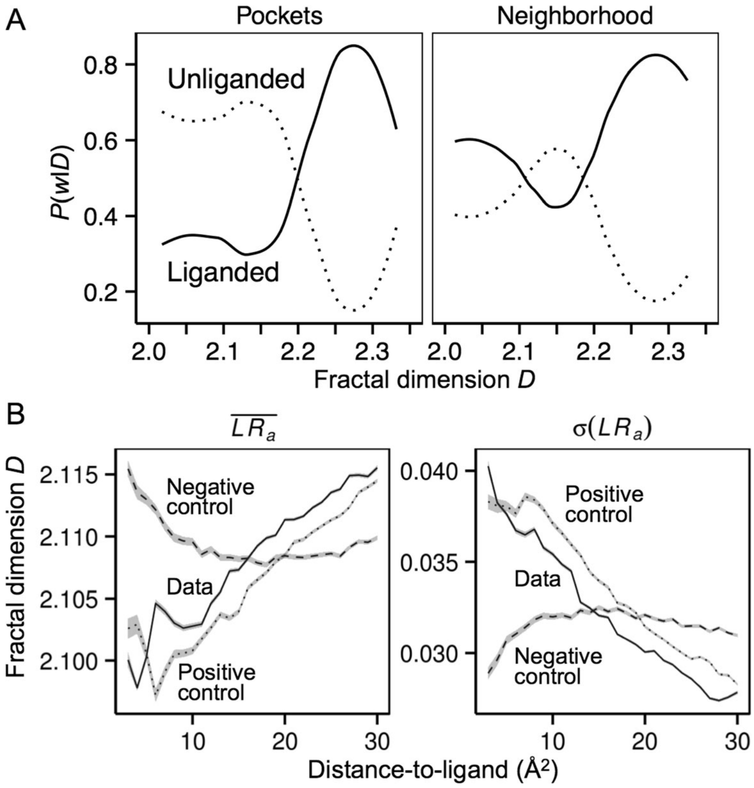

Figure 3. A) The class-conditional probability *P(w|D)* of a liganded or unliganded surface (*w*) given a certain observed fractal dimension (D). Pockets (left, $n=13{,}148$ surface cavities) were extracted using the PocketPicker method.[11] The classifier clearly distinguished known liganded pockets from unliganded pockets. In the atom-neighborhood context (right, $n=791{,}476$ surface patches), "Liganded" refers to protein surface patches with center atoms located ≤3 Å from the corresponding bound ligand. "Unliganded" refers to protein surface patches with center atoms located >3 Å away from a bound ligand. B) The distance-to-ligand dependency of the fractal dimension *D* for neighboring atoms, $LR_a(r)$ ($n=791{,}476$ surface patches; $r=6$ Å). The average values of all of the neighborhood means (left) and standard deviations σ (right) are presented as lines. The shaded areas give the standard error of the mean. Negative control: randomly selected atoms (dashed line). Positive control: the protein surface atom closest to the geometric center of the corresponding ligand (dotted line).

Table 1. Machine learning performance of the $LR_a(r)$ model assessed by stratified 10-fold cross-validation (*mean ± stddev.*). Abbreviations: $\mu=LR_a(r)$; $\sigma=\sigma(LR_a(r))$; s_a = surface patch area; NB: naïve Bayes classifier; NN: 25-nearest neighbor classifier; AUC: area under the curve; ROC: receiver operator characteristic.

Feature	Classifier	10-fold stratified cross-validation			
		AUC(ROC)	Sensitivity	Specificity	Balanced accuracy
D	NB	0.58 ± 0.02	0.34 ± 0.03	0.79 ± 0.03	0.56 ± 0.03
μ		0.61 ± 0.02	0.41 ± 0.03	0.78 ± 0.02	0.59 ± 0.03
σ		0.69 ± 0.02	0.48 ± 0.03	0.77 ± 0.02	0.62 ± 0.03
[μ; σ]		0.71 ± 0.02	0.53 ± 0.03	0.77 ± 0.02	0.65 ± 0.03
[μ; σ; s_a]		0.72 ± 0.02	0.55 ± 0.03	0.78 ± 0.03	0.66 ± 0.03
D	NN	0.57 ± 0.02	0.51 ± 0.03	0.59 ± 0.03	0.55 ± 0.03
μ		0.59 ± 0.02	0.52 ± 0.03	0.63 ± 0.03	0.57 ± 0.03
σ		0.66 ± 0.02	0.64 ± 0.03	0.60 ± 0.03	0.62 ± 0.03
[μ; σ]		0.69 ± 0.02	0.66 ± 0.03	0.64 ± 0.03	0.65 ± 0.03
[μ; σ; s_a]		0.73 ± 0.02	0.70 ± 0.02	0.66 ± 0.02	0.68 ± 0.03

classifier.[12] Among the methods tested, the naïve Bayesian algorithm[13] showed the highest specificity (0.78 ± 0.03) without sacrificing much of the sensitivity (0.55 ± 0.03) and was therefore the most restrictive (Table 1). We also applied a random forest[14] implementation as well as a multitude of voted combinations of three of the binary classifiers. All of the methods achieved balanced results, but they either underperformed compared with the naïve Bayesian and nearest-neighbor algorithms or were on par with them (Table 2). The best results were achieved when the features

Table 2. Machine learning performance of the $LR_a(r)$ model assessed by stratified 10-fold cross-validation (*mean*±*stddev.*). Abbreviations: $\mu = LR_a(r)$; $\sigma = \sigma(LR_a(r))$; s_a = surface patch area; NB: naïve Bayes classifier; NN: 25-nearest neighbor classifier; RF: 40-random forest classifier; VAP: voted average probability of MB, NN, and RF; VM: voted majority of MB, NN, and RF; VPP: voted product of probabilities of MB, NN, and RF; AUC: area under the curve; ROC: receiver operator characteristic.

Feature	Classifier	10-fold stratified cross-validation			
		AUC(ROC)	Sensitivity	Specificity	Balanced accuracy
D	RF	0.53±0.02	0.52±0.03	0.52±0.03	0.52±0.03
μ	RF	0.54±0.02	0.53±0.03	0.53±0.03	0.53±0.03
σ	RF	0.58±0.02	0.55±0.03	0.55±0.02	0.55±0.03
$[\mu;\sigma]$	RF	0.64±0.02	0.60±0.03	0.61±0.03	0.60±0.03
$[\mu;\sigma;s_a]$	RF	0.69±0.02	0.64±0.02	0.65±0.02	0.65±0.02
D	VAP	0.56±0.02	0.53±0.03	0.53±0.03	0.53±0.03
μ	VAP	0.58±0.03	0.55±0.03	0.55±0.03	0.55±0.03
σ	VAP	0.65±0.02	0.61±0.03	0.59±0.03	0.60±0.03
$[\mu;\sigma]$	VAP	0.69±0.02	0.63±0.03	0.66±0.03	0.64±0.03
$[\mu;\sigma;s_a]$	VAP	0.73±0.02	0.67±0.02	0.68±0.03	0.68±0.03
D	VM	0.55±0.02	0.45±0.03	0.67±0.03	0.56±0.03
μ	VM	0.58±0.02	0.48±0.03	0.68±0.03	0.58±0.03
σ	VM	0.62±0.02	0.58±0.03	0.67±0.03	0.62±0.03
$[\mu;\sigma]$	VM	0.64±0.02	0.68±0.03	0.64±0.03	0.65±0.03
$[\mu;\sigma;s_a]$	VM	0.67±0.02	0.66±0.03	0.70±0.03	0.68±0.03
D	VPP	0.54±0.02	0.52±0.03	0.53±0.03	0.53±0.03
μ	VPP	0.56±0.02	0.54±0.03	0.55±0.03	0.55±0.03
$;\sigma$	VPP	0.62±0.02	0.59±0.02	0.59±0.03	0.59±0.03
$[\mu;\sigma]$	VPP	0.69±0.02	0.62±0.03	0.68±0.03	0.64±0.03
$[\mu;\sigma;s_a]$	VPP	0.73±0.02	0.67±0.03	0.67±0.03	0.67±0.03

used as inputs for the program were a combination of $\overline{LR_a(r)}$, $\sigma(LR_a(r))$ and the patch surface area S_a.

We also computed predictions for an external dataset, which was compiled from the sc-PDB database.[15] This contained surface patches participating in known ligandable binding sites (n = 13,109,496 surface patches). In this dataset, our approach demonstrated a performance comparable to our internal dataset (balanced accuracy, naïve Bayesian algorithm: 0.64; nearest-neighbor algorithm: 0.64). Based on these encouraging results, we defined a *Lo*cal *R*oughness *I*ndicator (*LoRI*) model for predicting protein-ligand interaction hot-spots by training a naïve Bayesian classifier over the triple values of $[\overline{LR_a(r)}; \sigma(LR_a(r)); S_a]$. According to our *LoRI*, we expect approximately 22% of a protein's surface outside of the crystallographically known ligand binding sites to be targetable by small molecules.

The *LoRI* model performed stably in a sizable external test; therefore, we subjected the method to a series of retrospective and prospective surveys. We applied *LoRI* to concrete pharmaceutical targets not contained in our training data. To enable a visual inspection of the *LoRI* predictions, we projected the probabilities of ligandability that were calculated using *LoRI* onto the atoms contributing to the corresponding surface patch. We then colored atoms with higher probabilities (attractive) with warm colors, whereas atoms with lower probabilities (unattractive) were colored with colder colors.

The first blind test was performed on 4-diphosphocytidyl-2-C-methylerythritol synthetase (IspD), a cytidyltransferase in the non-mevalonate pathway for isoprenoid biosynthesis present in numerous prokaryota, algae, and protozoan parasites such as *Plasmodium falciparum*.[16,17] Because vertebrates synthesize isoprenoid precursors using a mevalonate pathway, all of the enzymes of the non-mevalonate pathway, including IspD, are interesting therapeutic targets. IspD is active as a dimer and catalyzes the conversion of 2-*C*-methyl-D-erythritol-4-phosphate (MEP) into 4-diphosphocytidyl-2-*C*-methyl-D-erythritol (CDP-ME).[18] *LoRI* successfully identified the active site of the co-crystal model, as indicated by the bright red coloring showing potential hot spots for ligand binding (Figure 4A).

The second blind test aimed to analyze a protein–protein interface because these macromolecular interaction sites are often flat and lack specific structural features.[19,20] We selected human interleukin-2 (IL-2), a cytokine signaling protein of the immune system. IL-2 binds to the α- and β-chains of the IL-2 receptor (IL-2R). IL-2 and the α-chain of IL-2R interact via a large surface patch.[21] Several complexes formed by IL-2 and small molecule inhibitors have been reported.[22] *LoRI* identified the inhibitor-binding site in the protein-protein interface based on the surface representation of IL-2 alone (Figure 4B). This result demonstrates that *LoRI* does not focus on cavities or buried residues but instead represents an independent view of molecular interaction sites.

To further challenge our approach, we applied it to a comparative (homology) model of HIV-1 protease subtype B. We computed *LoRI* values for each conformational snapshot of a 20 ns molecular dynamics simulation trajectory of the protease model (Figure 4C). *LoRI* identified the catalytic

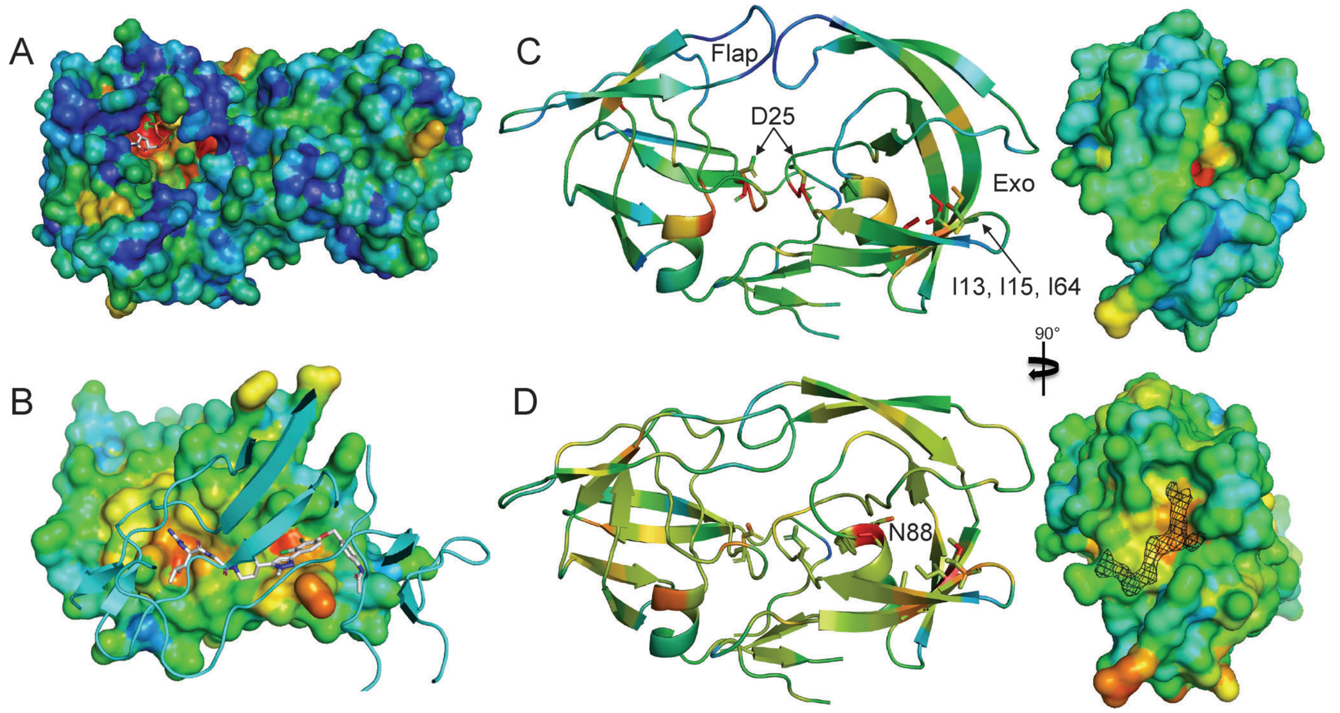

Figure 4. *LoRI* case studies. Warm colors indicate a higher predicted probability that the corresponding area is ligandable, while colder colors suggest the opposite. A) The IspD dimer in complex with CTP (PDB ID: 1I52).[16] B) IL-2 in complex with the α-chain of IL-2R (cartoon) (PDB ID: 1z92),[21] aligned and superimposed on IL-2 (surface) bound to a small molecule inhibitor (shown as sticks) (PDB ID: 1qvn).[35] In both test cases, *LoRI* was able to pinpoint the correct interaction location. C) A homology model of the HIV-1 protease subtype B colored by *LoRI* values averaged over the 20 ns MD simulation. D) A new crystal structure of the HIV-1 protease subtype B, which showed excess electron density in one of its two Exo-sites.

Table 3. The *LoRI* scores of residues in the active center and in the Exo-site of the HIV-1 protease. The values represent the highest atom ligandability scores per residue. The average per-residue scores are shown in parentheses.

Amino acid residue	Homology model	X-ray structure
Ile13	0.99 (0.75)	0.68 (0.53)
Ile15	0.74 (0.57)	0.86 (0.59)
Ile64	0.91 (0.71)	0.91 (0.61)
Asp25	1.00 (0.69)	0.50 (0.50)
Asn88	0.67 (0.56)	0.86 (0.69)
mean ± stddev.	0.43 ± 0.20	0.45 ± 0.12
min/max	0.11/1.00	0.12/0.91

Asp25 as ligandable, and predicted elevated ligandability probabilities for the residues flanking the elbow region (called the "Exo-site"), particularly one of the monomers (Table 3). We then crystallized HIV-1 protease subtype B and subjected this new crystal structure to *LoRI* analysis. Not only were the predictions of the homology model confirmed, we also observed pronounced electron density in one of the Exo-sites of the protease structure, thereby fully corroborating the computer-based prediction (Figure 4D). Similarly to the HIV-1 protease structure solved by Perryman et al. (PDB ID: 4e43),[23] this density could be attributed to the presence of bound acetate and glycerol.

3 Conclusions

In this study, we introduced the local fractal dimension of a molecular surface *D* as a unique and robust indicator of molecular interaction sites. The *LoRI* model represents a knowledge-based, alignment-free prediction approach that relies solely on surface roughness information. We verified this concept using blind tests and a prospective study, and our results suggest that *LoRI* has broad applicability to drug discovery and structural biology. We were also able to resolve previously discrepant interpretations of the meaning of molecular fractal dimension analysis[5,6] by showing that ligand binding sites possess higher *D* values than the rest of the protein's surface. Intriguingly, *LoRI* also provides a unifying concept for differently shaped protein surface areas that are attractive for protein–ligand interactions, including both deeply buried surface cavities as well as nearly planar protein–protein interaction sites. These findings suggest that local fluctuations in the fractal dimension of a surface of a putative binding site play an essential role in molecular recognition. For now, the physical nature of this phenomenon remains unknown, but we hope that our study will prompt further research aimed at solving this question. We are convinced that the fractal analysis of biological macromolecules holds much unexplored potential for structure analysis and drug design.

4 Materials and Methods

4.1 Fractal Scales and Density

Choosing the parameters for the calculation of the fractal dimension of a discretized non-strictly self-symmetrical natural object is the first step in its surface roughness assessment. There is a strict connection between the density of points of the discretized set, the fractal scales used for the calculation of D, and the D value itself (Figure 1). The choice of fractal scales may be summarized as follows: (i) choose exponential distances between the different scales; (ii) select the smallest scale greater than the resolution of the discretization, otherwise D will be severely overestimated (Figure 1B), and the largest scale to not exceed the size of the smallest object inspected, otherwise it will include a large amount of empty space which will result in underestimated D values. Following these rules, we chose fractal scales $\delta \in \{0.4, 0.8, 1.6, 3.2\}$ Å. The largest one is about the size of a small ligand or amino acid side-chain diameter. The next step in parameterizing the calculation of D is the selection of a suitable density of the points on the *SES* generated by MSMS[24] (Figure 1A). While calculating D for different densities between one and 20 ppÅ^2, we identified two intervals separated around 6 ppÅ^2, in which D behaved differently. The first interval harbors densities that are too low for the selected fractal scales, so that the hyperspheres for the smallest δ cannot reach any neighboring points, thus $c(X,\delta)=1/N$, which leads to overestimating D. This interval has a large dispersion compared to the other one (Figure 1A), but the orderings of the elements appear to be stable (Figure 1D), meaning that a switch in density inside the interval would not drastically change the overall ordering of the set. The same assumption also holds for the second interval containing densities greater than 6 ppÅ^2. It contains densities that are more favorable for a fractal dimension calculation, i.e. the production of stable regression lines, as well as the lower dispersion introduced by surface sampling effects. Naturally, a decrease in point density reduces the qualities of the discretization and results in a more crude approximation of D, but it also speeds up the computations, so we found a compromise at a density of 7 ppÅ^2. Note that choosing different fractal scales shifts the limit between the two intervals, depending on the choice of the lowest δ.

4.1.1 Local Roughness Model

Comprehensive Definition of the Model Parameters:

$$A = \{a(x,y,z) \mid a \text{ is an atom with positive SES}\}$$

$$A_a = \{b(x,y,z) \mid \|b-a\| < r;\ b \in A\}$$

$$S = \{s(x,y,z) \mid s \in \text{SES of the molecule}\}$$

$$S_a = \{s_a(x,y,z) \mid s \in \text{SES associated with } a\}$$

$$P_a = \{s_b \in S_b \mid \forall\, b \in A\}$$

$$D_{P_a} = \lim_{n\to\infty} \frac{\log C(X,\delta)}{\log \delta}$$

$$LR_a(r) = \{D_{Pa} \mid \|b-a\| < r\}$$

$$LoRI_a(r) = [\mu(LR_a(r)),\sigma(LR_a(r)),S_a]$$

4.2 Data Sets

The analyzed spatial molecular data relied exclusively on prior crystallographic surveys on protein-ligand complexes, DNA and RNA. The PDB ID lists are provided as Supplementary Information. Structures were compiled from different sources:

Protein-ligand-complexes. The protein-ligand complexes used in this survey were selected from two publicly available databases: (i) The PDBbind[25] (www.pdbbind.org) was accessed to create the *main set* of complexes, the results of which were verified on (ii) a subset of the sc-PDB,[15] termed the *verification set* for reference.

Main set. A subset of 604 protein-ligand complexes based on the PDBbind database was selected according to the following criteria: (i) Monomeric structures with predicted primary binding site volume not exceeding the volume of the corresponding ligand by more than 50%; (ii) structures for which the MSMS SES calculation failed due to software errors were excluded. The main set (Supplementary Listing 1) provided the following five sets of spatial data used in the survey at hand: (i) The proteins set contains spatial data gathered by stripping any atoms from the complexes that did not belong to a protein chain such as waters and ligands resulting in a set of 604 proteins; (ii) The ligands set contains the ligands of the active sites of the protein-ligand-complexes resulting in a set of 604 small molecules; (iii) The liganded pockets set contains the main pockets of the complexes resulting in a set of 604 active sites; (iv) The protein surface clefts set consists of pockets selected by PocketPicker[11] that are missing a ligand in the published crystallographic model; (v) The background surface patches set comprises of protein points subsets extracted using MSMS at uniformly distributed random positions on the surfaces of the 604 proteins.

External test set. A subset of 8639 protein-ligand complexes based on the sc-PDB database was selected according to the following criteria: (i) The complexes are not present in the main set; (ii) Very large structures (more than 2.35×10^5 surface points) and structures for which the

MSMS SES calculation failed due to software errors were excluded.

4.2.1 Pockets and Clefts

The positions of the pockets and unliganded clefts were identified by PocketPicker.[11] The surface patches for the pockets and clefts were excised using a hard proximity margin of 3 Å around the output of PocketPicker.

4.2.2 Random Surface Patches

For the sampling of the background surface patches 20 random atoms on the surface of each protein were selected. Each single surface patch was then defined by using spheres with 6 Å radius around each randomly selected surface atom.

4.2.3 Protein–Protein Interfaces

The PPI set was extracted from the freely available Dockground PPI database (dockground.bioinformatics.ku.edu/), benchmark 3.0.[26] It contains 131 complexes with both bound and one available unbound structures for each entry, as well as 102 complexes, each with both bound and both unbound structures available and provided. Thus, we were able to inspect 233 complexes with a grand total of 466 PPIs. Since we analyzed PPIs of known bound complexes, we used the corresponding complex partners for the definition of the interface's position, and a hard proximity margin of 3.5 Å. We chose this to be slightly higher than 3.0 Å, because at 3.0 Å the sampled patches for the PPI exhibited cut-off artifacts such as holes in the surfaces influencing the roughness estimation.

4.3 X-Ray Structure Analysis

HIV-1 protease, catalog number RH1P0001 was purchased from Biovendor, Heidelberg, Germany. The protein, delivered in 20 mM acetate buffer, pH 5.0 with 200 mM NaCl, 10 % (v/v) glycerol and 0.05 % 2-mercaptoethanol, was concentrated to a protein concentration of 3 mg/mL. 0.5 µL of a 200 mM HIV protease inhibitor (compound 7)[27] solution in DMSO was added to 80 µL of the concentrated HIV-1 protease solution. Because it turned out that the inhibitor was poorly soluble in the resulting protein solution, an additional 2 µL of a 50 mM Inhibitor solution in PEG400 was added to the last 20 µL of the protein solution. 0.1 µL of this latter protein solution was mixed with 0.1 µL reservoir solution of 0.1 M NaCitrate pH 3.5 and 3 M NaCl and equilibrated at 20 °C in a sitting drop setup. Small, bar-like crystals appeared after 1–3 days. For data collection, 20 % glycerol plus inhibitor was added to a crystal and the crystal was flash-frozen in liquid nitrogen. Data were collected on beam-line ID-29 at the European synchrotron radiation facility (ESRF) in Grenoble. Data processing was done with XDS[28] and scaling with the CCP4 program scala,[29,30] as implemented in the Global Phasing autoProc procedure.[31] The crystal diffracted to 2.28 Å with an overall Rmerge of 10.1 %. The space group was the same as of other HIV protease crystal structures. The crystal structure was solved by molecular replacement using the structure of HIV-1 protease (PDB ID: 5hpv)[32] as a model. Model building was done with *coot*[33] and the structure was refined with *Refmac*.[34] Data collection and refinement statistics are listed in Table 4. We found a peptide fragment fitted by Perryman et al. (PDB ID: 4e43)[23] in the active site of their 1.54 Å structure of "apo"HIV protease. In our case, the peptide appeared to be in a large part cleaved by the protease and only the P1-P4 residues were fitted. Weaker density indicated that partially also intact peptide was bound. Similarly to the 4e43 structure, an acetate ion and glycerol molecule were found bound in the exosite.

Table 4. Crystallographic data collection and refinement statistics. The highest resolution bin is given in brackets.

Data collection		Refinement	
Space group	$P2_12_12$	Protein atoms	1555
Cell dimensions:		Acetate ions	1
a,b,c (Å)	58.24, 86.50, 45.92	Glycerol molecules	3
a,b,g (°)	90.00, 90.00, 90.00	DMSO molecules	3
Resolution (Å)	86.50, −2.28 (2.41–2.28)	Water molecules	290
$\langle I\rangle/\sigma\langle I\rangle$	14.7 (4.2)	Other atoms	32
Observed reflections	70332 (10394)	Resolution (Å)	48.31–2.28 (2.34–2.28)
Rmerge	0.101 (0.479)	Rwork (%)	16.0 (22.7)
Rmeas	0.121 (0.597)	Rfree (%)	24.4 (40.0)
Completeness (%)	99.9 (99.8)	Average B-factors (Å^2):	
Redundancy	6.4 (6.6)	Protein	20.93
		Water	40.26
		Other	48.28
		rmsd bond lengths (Å)	0.015
		rmsd bond angles (°)	1.75

Conflict of Interest Statement

G. S. is a co-founder of inSili.com LLC, Zürich, and a consultant in the pharmaceutical and chemical industry.

Acknowledgements

We thank Dr. *M. Rupp* for the stimulating discussion on computerized learning. This research was financially supported *by ETH Zürich* and by *Sanofi-Aventis Deutschland*.

References

[1] a) S. Leis, S. Schneider, M. Zacharias, *Curr. Med. Chem.* **2010**, *17*, 1550–1562; b) S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, B. O. Villoutreix, *Drug Discov. Today* **2010**, *15*, 656–667; c) E. B. Fauman, B. K. Rai, E. S. Huang, *Curr. Opin. Chem. Biol.* **2011**, *15*, 463–468.

[2] a) H. Li, V. Kasam, C. S. Tautermann, D. Seeliger, N. Vaidehi, *J. Chem. Inf. Model.* **2014**, *54*, 1391–1400; b) J. Desaphy, K. Azdimousa, E. Kellenberger, D. Rognan, *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299; c) T. Geppert, B. Hoy, S. Wessler, G. Schneider, *Chem. Biol.* **2011**, *18*, 344–353; d) P. Schmidtke, X. Barril, *J. Med. Chem.* **2010**, *53*, 5858–5867; e) T. A. Halgren, *J. Chem. Inf. Model.* **2009**, *49*, 377–389; f) M. Weisel, E. Proschak, J. M. Kriegl, G. Schneider, *Proteomics* **2009**, *9*, 451–459.

[3] a) J. Li, P. Mach, P. Koehl, *Comput. Struct. Biotechnol. J.* **2013**, *8*, e201309001; b) M. Stahl, C. Taroni, G. Schneider, *Protein Eng.* **2000**, *13*, 83–88; c) M. Weisel, J. M. Kriegl, G. Schneider, *ChemBioChem* **2010**, *11*, 556–563.

[4] J. C. Russ, *Fractal Surfaces*, Springer, New York, **1994**.

[5] M. Lewis, D. C. Rees, *Science* **1985**, *230*, 1163–1165.

[6] F. K. Pettit, J. U. Bowie, *J. Mol. Biol.* **1999**, 285, 1377–1382.

[7] M. F. Sanner, A. J. Olson, J. C. Spehner, *Biopolymers* **1996**, *38*, 305–320.

[8] P. Grassberger, I. Procaccia, *Phys. Rev. Lett.* **1983**, *50*, 346–349.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.

[10] Z. Shentu, M. Al Hasan, C. Bystroff, M. J. Zaki, *Proteins* **2008**, *70*, 1056–1073.

[11] M. Weisel, E. Proschak, G. Schneider, *Chem. Cent. J.* **2007**, *1*, 7.

[12] E. Fix, J. L. Hodges, *USAF School of Aviation Medicine* **1951**, 261–279.

[13] N. Friedman, D. Geiger, M. Goldszmidt, *Machine Learning* **1997**, *29*, 131–163.

[14] L. Breiman, *Machine Learning* **2001**, *45*, 5–32.

[15] J. Meslamani, D. Rognan, E. Kellenberger, *Bioinformatics* **2011**, *27*, 1324–1326.

[16] S. B. Richard, M. E. Bowman, W. Kwiatkowski, I. Kang, C. Chow, A. M. Lillo, D. E. Cane, J. P. Noel, *Nat. Struct. Biol.* **2001**, *8*, 641–648.

[17] H. K. Lichtenthaler, J. Zeidler, J. Schwender, C. Müller, *Z. Naturforsch.* **2000**, *C* 55, 305–313.

[18] F. Rohdich, J. Wungsintaweekul, M. Fellermeier, S. Sagner, S. Herz, K. Kis, W. Eisenreich, A. Bacher, M. H. Zenk, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11758–11763.

[19] L. L. Conte, C. Chothia, J. Janin, *J. Mol. Biol.* **1999**, *285*, 2177–2198.

[20] W. E. Stites, *Chem. Rev.* **1997**, *97*, 1233–1250.

[21] M. Rickert, X. Wang, M. J. Boulanger, N. Goriatcheva, K. C. Garcia, *Science* **2005**, *308*, 1477–1480.

[22] R. Bourgeas, M.-J. Basse, X. Morelli, P. Roche, *PLoS One* **2010**, *5*, e9598.

[23] A. L. Perryman, Q. Zhan, H. H. Soutter, R. Rosenfeld, D. E. McRee, A. J. Olson, J. E. Elder, C. D. Stout, *Chem. Biol. Drug Des.* **2010**, *75*, 257–268.

[24] M. F. Sanner, A. J. Olson, J. C. Spehner, *Biopolymers* **1996**, *38*, 305–20.

[25] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, *J. Med. Chem.* **2005**, *48*, 4111–4119.

[26] Y. Gao, D. Douguet, A. Tovchigrechko, I. A. Vakser, *Proteins* **2007**, *69*, 845–851.

[27] J. Kunze, N. Todoroff, P. Schneider, T. Rodrigues, T. Geppert, F. Reisen, H. Schreuder, J. Saas, G. Hessler, K.-H. Baringhaus, G. Schneider, *J. Chem. Inf. Model.* **2014**, *54*, 987–991.

[28] W. Kabsch, *Acta Crystallogr.* **2010**, *D66*, 125–132.

[29] Collaborative Computational Project, Number 4, *Acta Crystallogr.* **1994**, *D50*, 760–763.

[30] P. R. Evans, *Acta Crystallogr.* **2006**, *D62*, 72–82.

[31] C. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, G. Bricogne, *Acta Crystallogr.* **2011**, *D67*, 293–302.

[32] P. M. Fitzgerald, B. M. McKeever, J. F. VanMiddlesworth, J. P. Springer, J. C. Heimbach, C. T. Leu, W. K. Herber, R. A. Dixon, P. L. Darke, *J. Biol. Chem.* **1990**, *265*, 14209–14219.

[33] P. Emsley, K. Cowtan, *Acta Crystallogr.* **2004**, *D60*, 2126–2132.

[34] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, *Acta Crystallogr.* **2011**, *D67*, 355–367.

[35] C. D. Thanos, W. L. DeLano, J. A. Wells, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 15422–15427.

Received: June 27, 2014
Accepted: June 30, 2014
Published online: September 2, 2014