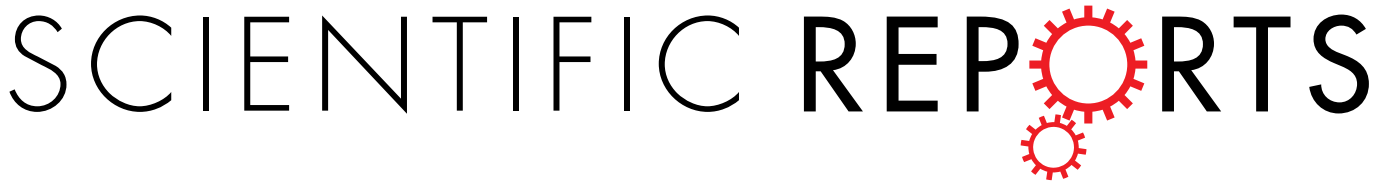

OPEN

Cutoff lensing: predicting catalytic sites in enzymes

Simon Aubailly & Francesco Piazza

Received: 28 June 2015
Accepted: 10 September 2015
Published: 08 October 2015

Predicting function-related amino acids in proteins with unknown function or unknown allosteric binding sites in drug-targeted proteins is a task of paramount importance in molecular biomedicine. In this paper we introduce a simple, light and computationally inexpensive structure-based method to identify catalytic sites in enzymes. Our method, termed *cutoff lensing*, is a general procedure consisting in letting the cutoff used to build an elastic network model increase to large values. A validation of our method against a large database of annotated enzymes shows that optimal values of the cutoff exist such that three different structure-based indicators allow one to recover a maximum of the known catalytic sites. Interestingly, we find that the larger the structures the greater the predictive power afforded by our method. Possible ways to combine the three indicators into a single figure of merit and into a specific sequential analysis are suggested and discussed with reference to the classic case of HIV-protease. Our method could be used as a complement to other sequence- and/or structure-based methods to narrow the results of large-scale screenings.

With the rapid development and refinement of experimental techniques for protein structure determination at high resolution, predicting functional sites is a major issue in modern molecular biology in many protein families[1–7].

The swiftly growing amount of structural and sequence data poses big challenges and offers great opportunities to test automated prediction algorithms and platforms. Several approaches have been used to identify critical function-related sites (sometimes referred to as *hotspots*) in proteins. Most of these methods imply structural and/or sequence conservation information[8–14].

Purely sequence conservation approaches use phylogenetic information, relying on the idea that functional sites are conserved during evolution. Typically, such algorithms proceed through the alignment of a great number of different sequences and the ensuing computation of different conservation scores[7,15–18]. Other approaches can be found in the literature, typically combining sequence-related information with structural data to achieve higher prediction rates[19–21].

Among the structure-based algorithms developed to identify and predict function-related sites in proteins, an appealing and promising class is that of coarse-grained (CG)[22,23] approaches based on elastic-network models (ENM)[24–29]. The ENM[30] and its CG versions[31,32] are light and computationally inexpensive tools that have proved tremendously effective in dissecting function-related vibrational patterns in proteins, both embodied in low-frequency collective normal modes[33–37] and, more subtly, related to high-frequency localized vibrations[28,38–41].

Often, graph-theoretical tools have been employed in combination with ENM-related approaches[42–50] to identify hotspots and binding interfaces. In these methods, a protein structure is mapped onto a network by means of some rule. In one simple CG example, nodes may represent amino acids while edges embody pair-wise interactions that can be obtained either from the study of equilibrium structures[47,48] or from molecular dynamics (MD) simulations[51]. Typical graph-theoretical measures employed for such analyses include connectivity[24,48], different measures of centrality[16,45,47,52–54], betweenness and cluster coefficient[16].

It is clear from the above discussion that a successful strategy to predict functional sites in proteins has to rely on a composite approach, combining information from sequence conservation with structure-based analyses. In turn, the latter should combine different indicators, related to the

Université d'Orléans, Centre de Biophysique Moléculaire, CNRS-UPR4301, Rue C. Sadron, 45071, Orléans, France. Correspondence and requests for materials should be addressed to F.P. (email: Francesco.Piazza@cnrs-orleans.fr)

physical-chemical properties of amino acid environments and to patterns of chemical and topological connectivity.

In this paper we focus on the prediction of catalytic sites in enzymes based on an original ENM-based strategy. Atomistic approaches devised to identify residues involved in catalysis in enzymes are not new[55]. More recently, approaches specifically relying on sophisticated electrostatic calculations have been introduced[56,57]. Conversely, coarse-grained models have been relatively less exploited to solve this specific problem[29,58,59]. Yet, ENM-based tools are light (they can be applied to large databases of structures) and can be readily extended to perform all-residue searches in many structures. Moreover, CG topology-based methods have the advantage to strip the structure of most chemical details so as to bring to the surface purely topological features. This appears particularly important in the case of enzymes, as often sites that are involved in the catalytic action are intriguingly found far from the annotated catalytic sites[60,61].

Our method combines three different indicators, two graph-theoretical measures with an original scale of local *stiffness* in a method that we termed *cutoff lensing*. The main idea is that catalytic sites can be spotlighted by employing elastic network models whose connectivity is increased beyond currently employed values. In ENMs, a spring is stretched between all pairs of residues that are separated in the equilibrium structure by a distance less than a specified cutoff length R_c. Typically employed values for protein models coarse-grained at the level of amino acids vary in the 10–13 Å range[32,62–65], even if values greater than 13 Å have also been considered episodically[66–68]. In principle, larger values of the cutoff are unphysical, as the connectivity graph becomes nearly fully connected as R_c attains a value comparable with the protein size. Nevertheless, we have found that specific, function-related sites can be singled out in such regimes by using indicators associated with topological and structural measures of connectedness and stiffness. Remarkably, a scan of increasing values of the cutoff shows that there exists an optimum range where our structural indicators are the most sensitive in detecting catalytic sites known from experiments. This *lensing* effect can thus be used to predict the location of functional sites in unannotated proteins.

The paper is organized as follows. In the next section we provide the description of the cutoff lensing method and introduce three structure-based indicators. In the following section we check the predictive power of our indicators against the pool of annotated catalytic sites in a large database of enzymes. Finally, we discuss our results and provide a working summary of our method.

Methods

We model a given protein consisting of N amino acids as an ensemble of N fictitious particles occupying the equilibrium positions of the corresponding α-carbons, as found in the experimental structure. All particles have the same mass M, which we set equal to the average amino acid mass, $M = 120$ Da (as the fictitious particles occupy the equilibrium positions of amino acids, *i.e.* are located on the corresponding α-carbons, we will use the words particles and (amino acid) residues interchangeably). Each particle interacts with its neighbours, as specified by the cutoff distance R_c. More precisely, residues i and j interact if $|\boldsymbol{R}_i - \boldsymbol{R}_j| \leq R_c$, where $\boldsymbol{R}_i$ denotes the position vector of the i-th residue in the equilibrium structure. Let $\boldsymbol{r}_i$ denote the instantaneous position vector of the i-th residue. The system potential energy reads

$$\mathcal{U} = \frac{1}{2} k_2 \sum_{i>j} c_{ij} \left(r_{ij} - R_{ij} \right)^2 \tag{1}$$

where $r_{ij} = |\boldsymbol{r}_{ij}| = |\boldsymbol{r}_i - \boldsymbol{r}_j|$, $R_{ij} = |\boldsymbol{R}_{ij}| = |\boldsymbol{R}_i - \boldsymbol{R}_j|$ are the inter-particle instantaneous and equilibrium distances, respectively. Interacting pairs are specified through the contact matrix

$$c_{ij} = \begin{cases} \theta\left(R_c - R_{ij}\right) & i \neq j \\ 0 & i = j \end{cases} \tag{2}$$

$\theta(x)$ denoting the Heaviside step function. Eq. (1) amounts to building a network of Hookean springs joining pairs of residues separated by a distance smaller than R_c. Normal modes (NM) are computed by diagonalizing the (mass-weighted) Hessian matrix,

$$\widetilde{\mathbb{H}}_{ij}^{\alpha\beta} = \frac{1}{M} \frac{\partial^2 \mathcal{U}}{\partial r_{i\alpha} \partial r_{j\beta}} \bigg|_{r=R} \tag{3}$$

which gives $3N-6$ normal modes $\boldsymbol{\xi}_i^k$, $k = 7, 8, \ldots, 3N$ with non-zero eigenvalues ω_k^2. Here, greek indexes indicate Cartesian spatial directions. It is straightforward to show from Eq. (1) that

$$\widetilde{\mathbb{H}}_{mj}^{\alpha\beta} = -\omega_0^2 \left(c_{mj} \hat{R}_{mj}^{\alpha} \hat{R}_{mj}^{\beta} - \delta_{jm} \sum_{k \neq m} c_{mk} \hat{R}_{mk}^{\alpha} \hat{R}_{mk}^{\beta} \right) \tag{4}$$

where δ_{jm} is the Kronecker symbol, $\hat{\boldsymbol{R}}_{ij} = \boldsymbol{R}_{ij}/R_{ij}$ and $\omega_0 = \sqrt{k_2/M}$. Following previous studies, we fix $k_2 = 5$ kcal/mol/Å^2 [63].

Constructing structural hotspot indicators. The basic idea of our method rests on the evidence reported by several studies that hotspot/functional sites in proteins are generally found in stiff/rigid regions[59,69–71]. Analogously, it has been shown that functional residues tend to move independently from the rest of the structure, involving high-frequency localized vibrations (the stiffer the bonds, the higher the frequency)[62,63,72,73].

Structural rigidity can be gauged by many indicators, that assess the different *flavours* associated with it. The simplest and more intuitive method, albeit unsuitable for automated screening of large structure databases, would be to measure fluctuations directly via MD simulations, such as in ref. 74. Alternatively, but more indirectly, rigidity can be related to the local number of neighbours in the protein connectivity graph. A series of recent studies has demonstrated a rather surprising agreement between the location of catalytic sites in enzymes and the localization patterns of nonlinear vibrational modes known as discrete breathers (DB)[38,39]. Such observations have been rationalized in terms of a *spectral measure of local stiffness*, based on the localization properties of high-frequency normal modes[62].

Here we introduce an original method based on a blend of suitable structural indicators combined with *cutoff lensing*, *i.e.* an analysis where the cutoff R_c is let increase beyond physically realistic values. The key feature of this method is a selective sharpening of the predictive power of our indicators at specific intermediate values of R_c.

A *spectral* stiffness measure can be computed by looking at the contribution of a reduced set of high-frequency NMs, $\mathcal{S}_{hf}$, to atomic fluctuations, that is

$$\chi_i = \sum_{k\in\mathcal{S}_{hf}} |\boldsymbol{\xi}_i^k|^2 \tag{5}$$

In the following we shall consider the last five high-frequency NMs, *i.e.* $\mathcal{S}_{hf} \equiv [3N-4,\ 3N]$. The rationale behind Eq. (5) comes from the observation that fast normal modes tend to be localized at hotspot sites[62], *i.e.* sites that act as efficient energy storage and accumulation centers, typically flagging highly connected and buried regions. Along the same lines, fast modes have also been demonstrated to identify stability cores of proteins[41], adding to the meaningfulness of definition (5). Typically, in residue-based coarse-grained ENMs the last high-frequency NMs are localized around one, two sites at most. If one considers an average number of catalytic sites per enzyme around 5 (it is 2.5 in the Catalytic Site Atlas (CSA)[75]), it appears that the minimum number of high-frequency NMs to include in the definition (5) is five (adding a few more NMs does not change appreciably our results. Adding more results in useless blurred patterns).

Following a similar rationale, we shall also consider indicators referring to the connectivity graph, notably the local connectivity,

$$c_i = \sum_j c_{ij} \tag{6}$$

and, as already done by other authors[16,45,47,52–54], the closeness centrality, defined as

$$CC_i = \left(\sum_j \ell_{ij}\right)^{-1} \tag{7}$$

where ℓ_{ij} is the shortest path (in units of edge number) between nodes i and j over the connectivity graph.

The three above-defined indicators can be regarded as supplying different measures of *stiffness*. While χ_i gauges the *vibrational* stiffness of a given residue, *i.e.* its propensity to vibrate at high frequency with a space-localized pattern, c_i and CC_i exquisitely quantify the *topological* stiffness, in the sense of number of outgoing bonds (c_i) or shortest paths between two given locations flowing through i (CC_i).

As a general rule, the raw measures of χ_i, c_i and CC_i result in rather rugged and irregular patterns with many peaks and troughs for a given protein sequence. Our goal is to extract from such patterns the most relevant peaks as flags for potentially functional sites. To this aim, we apply a high-pass *filtering* procedure, by keeping for a given indicator pattern only the values above a specified number of standard deviations (computed over the whole sequence). More precisely,

$$\widetilde{\mathcal{I}}_j = \begin{cases} \mathcal{I}_j & \text{for } \mathcal{I}_j > N_{\mathcal{I}}\sigma_{\mathcal{I}} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $\mathcal{I}_i$ is the considered indicator, $\sigma_{\mathcal{I}} = \sqrt{\sum_j (\mathcal{I}_j - \langle\mathcal{I}\rangle)^2/(N-1)}$ its standard deviation and $N_{\mathcal{I}}$ an indicator-dependent high-pass threshold. In our analysis, we fixed $N_{\mathcal{I}} = 3$ for $\mathcal{I} = \chi$, CC and $N_{\mathcal{I}} = 5$ for $\mathcal{I} = c$. Our final site predictions are then obtained as the locations flagged by the peaks that survive in the pattern after the high-pass filtering. In the case of CC, the patterns showed overly rugged profiles (see Fig. 1), which resulted in a large number of close, quasi-degenerate peaks after the high-pass

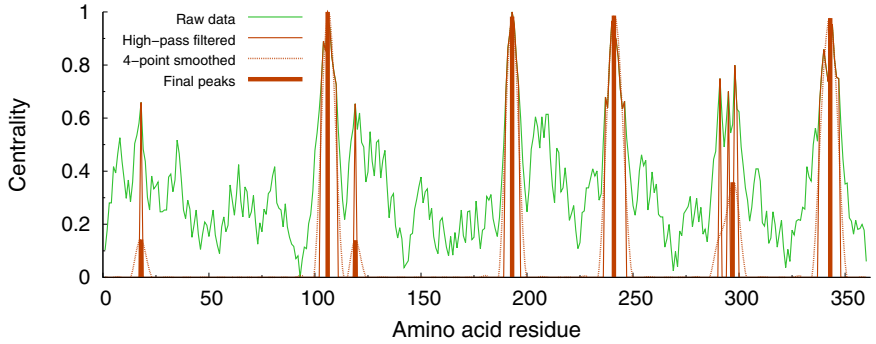

Figure 1. Illustration of the computation of the reduced closeness centrality indicator through the different sequential steps described in the text. The patterns are normalized to the maximum value occurring in the sequence. The final peaks flag the potentially functional sites. The calculations refer to Arginin Glycineaminotransferase (PDB code 1JDW).

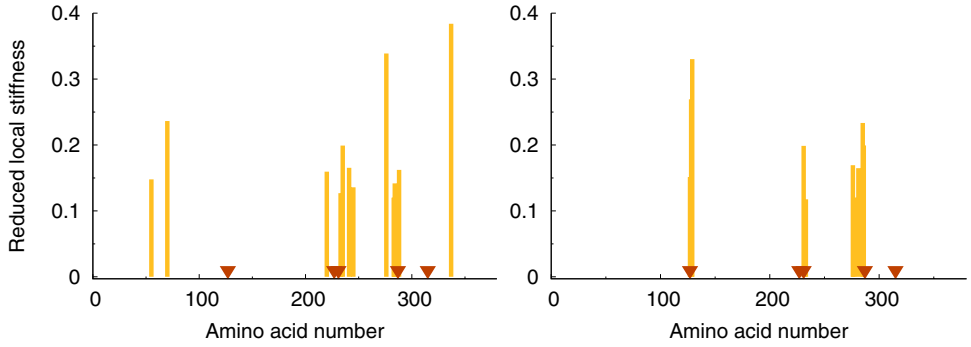

Figure 2. Illustration of the *cutoff lensing* effect. Plot of the reduced stiffness pattern $\widetilde{\chi}$, Eq. (8), for Arginin Kinase (PDB code 1BG0). Cutoff $R_c = 10$ Å (left) and $R_c = 20$ Å (right). The known catalytic sites are indicated by dark triangles. Note the disappearance of some *irrelevant* peaks and the appearance of a peak at one of the catalytic sites in going from $R_c = 10$ Å to $R_c = 20$ Å.

filtering. Accordingly, in order to eliminate the degeneracy associated with multiple-peak structures, we applied a 4-point smoothing procedure[76] to the filtered patterns, so as to automatically make the excessively degenerate structures coalesce in one single-peak prediction. The whole procedure is illustrated in Fig. 1.

Results: the cutoff lensing effect

Our idea is to inspect reduced (filtered) patterns of local spectral and topological stiffnesses in search for hot spots. One of such patterns is reported in Fig. 2 for two values of the cutoff parameter R_c used to construct the elastic network (see again Eq. (2)). Interestingly, one may easily remark that there is a correspondence between the location of known catalytic sites and stiffness peaks. This finding agrees with observations made by other authors along the same lines[29,59]. However, if we now repeat the same analysis with a higher (even if less physical value of R_c), the surprising consequence is that the reduced pattern is sharpened down to a handful of peaks, which seem to much better pinpoint the known functional sites. Note that the observed sharpening of the predictive power implies both the *evaporation* and the *relocation* of some peaks. We term this effect altogether *cutoff lensing*.

The logical questions to ask in view of such findings are (i) whether these effects also characterize the other indicators and (ii) whether there exists an optimum value of R_c, corresponding to the maximum overlap between (generalized) stiffness peaks and catalytic sites, beyond which the patterns get blurred again and one correspondingly loses predictive power. The latter possibility, in particular, seems highly realistic, as one expects sites to be no longer distinguishable (with respect to whatever measure) in nearly fully connected networks.

The results reported in Fig. 3 for a given enzyme seem to reply to the first question in the affirmative: intermediate values of the cutoff appear to be associated with increased predictive power. The reduced connectivity $\check{c}_i$ and spectral stiffness $\widetilde{\chi}_i$ profiles suggest that intermediate values of R_c yield a better match between the peaks of the indicator patterns and the annotated sites. The centrality, on the contrary,

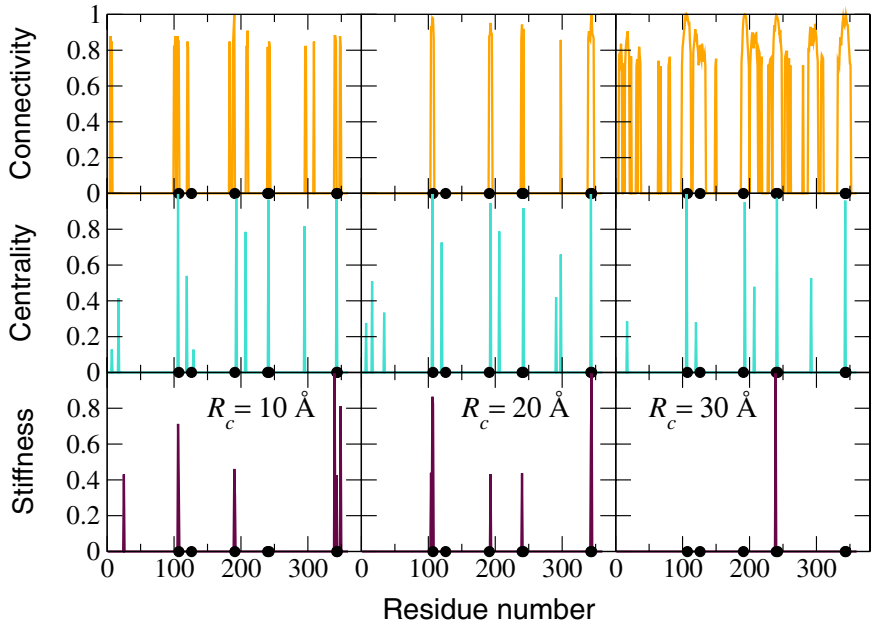

Figure 3. Reduced and normalized connectivity, closeness centrality and stiffness patterns computed according to the prescription (8) for Arginin Glycineaminotransferase (PDB code 1JDW) for different values of the cutoff R_C. The annotated catalytic sites are indicated by black filled circles.

provides a good match but seems at the same time rather insensitive to changes in the cutoff. It is important to observe that the number of peaks N_p is not constant as a function of R_c. Of course, this information has to be included in the picture if we want to provide a statistical assessment of the predictive power of our indicators as a function of R_c. On the one hand, N_p is expected to increase at high values of R_c for the connectivity and centrality measures, while it seems that stiffness patterns display less and less peaks as the cutoff is made larger.

In order to shed further light on the above-described findings and proceed to a statistical assessment of the ability of our indicators to spotlight function-related sites, we have analyzed a pool of 835 enzyme structures from the Catalytic Site Atlas[75]. The CSA is a major resource in the field of structural biology, and provides up-to-date catalytic residue annotation for enzymes in the Protein Data Bank based on experimental structural data. The results of our ensemble analysis are reported in Fig. 4. For each indicator, we have calculated the fraction of catalytic sites that are found within a prescribed distance Δn (in units of residues) along the sequence from a peak. For example, the curves at $\Delta n = 0$ indicate the fraction of catalytic sites that coincide with a peak for a given indicator.

A number of interesting observations can be made by inspecting Fig. 4. The reduced connectivity $\tilde{c}$ increases its predictive power at increasing values of the cutoff. However, this is a trivial consequence of the fact the number of peaks also increases as the systems become more and more connected (top right panel). Therefore, the connectivity does not appear to provide a particularly insightful spotlighting tool. On the contrary, the reduced centrality $\widetilde{CC}$ provides a comparatively more sensitive detection tool, with up to half of the whole pool of catalytic sites found at a separation of at most one residue (along the sequence) from a $\widetilde{CC}$ peak. Furthermore, it is seen that the predictive power of this indicator is almost insensitive to the number of peaks, which increases of course as the structures become more and more connected (middle right panel). Interestingly, the average number of peaks in the *CC* patterns displays a minimum (around $R_c = 28$ Å), which suggests that at this value of the cutoff the *reliability* of the observed predictive power of centrality is maximum.

Of the three indicators, the reduced stiffness $\widetilde{\chi}$ displays the most interesting behavior. The fraction of predicted sites shows a maximum at intermediate cutoff values (around 20 Å), with up to 30% of the known catalytic sites recovered at a distance of one amino acid from a peak of reduced stiffness. Interestingly, the number of such peaks decreases towards a nearly constant value as the cutoff is increased. Most remarkably, the maximum of predictive power clearly falls in a regime where the number of peaks has attained its minimum asymptotic value, which means that the statistical significance of the prediction at the maximum is also maximum. To make this observation more quantitative, one may introduce an intuitive measure of *reliability* defined as the fraction of predicted sites divided by the number of peaks found at each value of the cutoff. This is illustrated in Fig. 5 (left panel). It is clear that the predictions made from the reduced stiffness patterns correspond to a maximum of reliability at the intermediate cutoff $R_c \simeq 22$ Å. This suggests that the cutoff lensing effect can be effectively employed to predict the location of catalytic sites or to

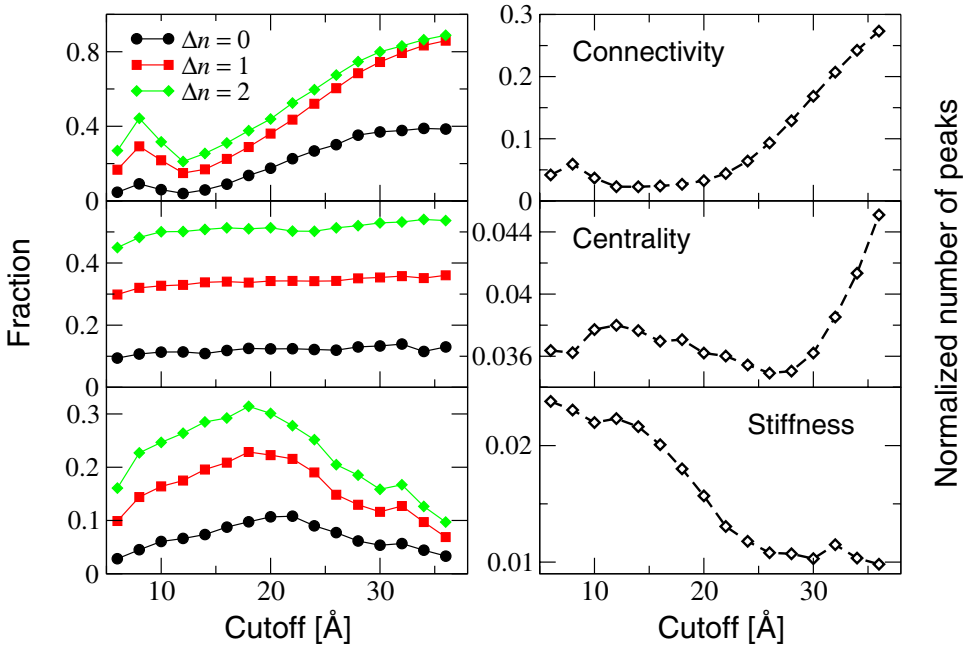

Figure 4. Left panels: fraction of catalytic residues within Δn sites from the nearest peak versus cutoff, as computed over the ensemble of enzymes from the CSA. Right panels: average peak fraction (number of peaks divided by number of residues) computed over the whole database versus cutoff.

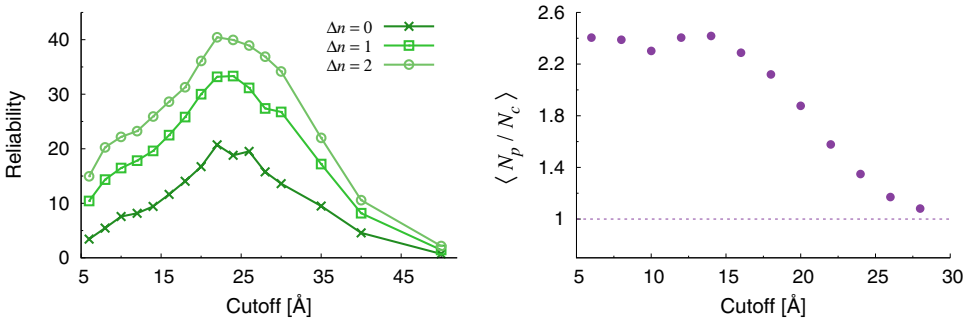

Figure 5. Left panel: reliability of the predictive power of reduced stiffness patterns as a function of the cutoff R_C (arbitrary units). The reliability is defined as the fraction of predicted catalytic sites (within Δn amino acids along the sequence) divided by the fraction of stiffness peaks (number of peaks per amino acid). Right panel: Average number of peaks in the reduced stiffness patterns per catalytic site.

substantiate the predictions made by means of other methods based on different arguments. This is also confirmed by the observation that the highest number of predicted sites and maximum reliability corresponds to roughly one stiffness peak per catalytic site (see right panel in Fig. 5). This suggests that the optimality condition of maximum predictive power is achieved with the least number of unassociated peaks, *i.e.* under conditions of highly reduced redundancy.

It is interesting to note that the cutoff value associated with the maximum in the fraction of sites predicted by the reduced stiffness patterns increases with the size of the protein, and so does the fraction itself at the maximum. This is illustrated in Fig. 6, where we show the results of our computations performed by grouping the enzymes in three different size classes. It is clear that our algorithm is much more effective for proteins of large sizes. This remarkable finding is not restricted to stiffness patterns. In general, the fraction of catalytic sites recovered by reduced closeness and connectivity profiles is greater for enzymes of larger sizes (see supplementary material).

Discussion, Conclusions and Perspectives

In this paper we have investigated the ability of different structure-related indicators to pinpoint the location of known catalytic sites in a large number of enzyme structures in the framework of the elastic network model. More precisely, we defined reduced *peak patterns* of (i) local connectivity, (ii) closeness

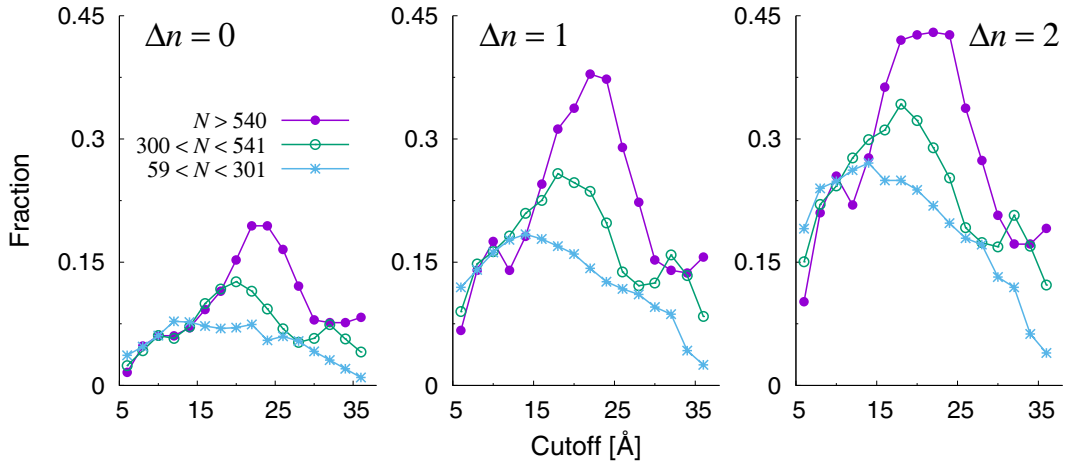

Figure 6. Fraction of catalytic sites within Δn sites from the nearest peak of the stiffness reduced patterns computed over three different size classes in the CSA database.

centrality and (iii) structural stiffness, where the peaks retained along the protein sequence are assumed to flag potentially interesting sites. Our method is general and computationally inexpensive (see supplementary material for a benchmark test).

Our analysis shows that all three considered indicators display a considerable predictive power (up to 50% of the catalytic sites recovered within a distance of two amino acids along the sequence), when the computed peak structures are compared with the location of annotated catalytic sites in a large database of enzymes (the Catalytic Site Atlas[75]). This suggests that the three indicators can be employed in some suitable combination/sequence to make predictions in unannotated enzyme structures.

In order to find the optimal procedure to combine the three indicators, we have investigated their behavior as a function of the cutoff R_c used to construct the elastic networks, while monitoring in parallel the number of peaks per amino acid present in the indicator patterns. We have termed this procedure *cutoff lensing*. This analysis has revealed that optimal values of the cutoff exist in all cases. For connectivity, the fraction of known catalytic sites recovered trivially (and uninformatively) increases with the cutoff, as the number of high-connectivity peaks retained also increases. For this reason, we argue that the optimal cutoff corresponds to the highest predictive power corresponding to the least number of peaks per amino acid (about 40% of the catalytic sites recovered within a distance of two amino acids along the sequence), which means $R_c \approx 20$ Å. By contrast, somewhat surprisingly, centrality patterns display nearly cutoff-invariant predictive power. However, the specific number of peaks displays a minimum around $R_c = 28$ Å. Therefore, we conclude that $R_c = 28$ Å can be taken as the optimality condition, reflecting the idea that for equal fractions of recovered catalytic sites the most reliable prediction is the one made with the least number of peaks.

The study of reduced stiffness patterns has led us to uncover an interesting effect, that we termed *cutoff lensing*: when the cutoff is increased, the fraction of catalytic sites spotlighted by the stiffness peak patterns displays a maximum at around $R_c = 20$ Å. Remarkably, this is achieved with a minimum degree of redundancy, as the number of peaks in the patterns (pointing to potentially interesting sites) is a minimum for $R_c > 22$ Å, while at the same time the average number of peaks per catalytic sites is about 1 on average in this range of cutoff values. We conclude that $R_c \simeq 22$ Å is the value of choice for predictions of catalytic sites made through stiffness patterns.

Remarkably, we found that the fraction of catalytic sites recovered by our indicators at the optimal cutoff is larger the larger the protein (see again Fig. 6 and also the Supplementary Material). Connectivity patterns are an exception, as at the optimal cutoff the fraction of catalytic sites recovered is nearly the same independently of the size of the enzymes.

A sequential computation of optimal indicators to make predictions, combined scores to assess their confidence level. It is interesting to inquire whether it is possible to combine the three indicators computed at their individually optimal cutoff values in some *globally optimal* manner and what would be the meaning of such combination. The simplest operation to do is to add up the information carried by the three figures of merit, weighted by the number of peaks displayed by the corresponding patterns. This leads to introducing the following global score,

$$S_i = \frac{1}{3}[\sigma_i^{\chi} + \sigma_i^{CC} + \sigma_i^{c}] \tag{9}$$

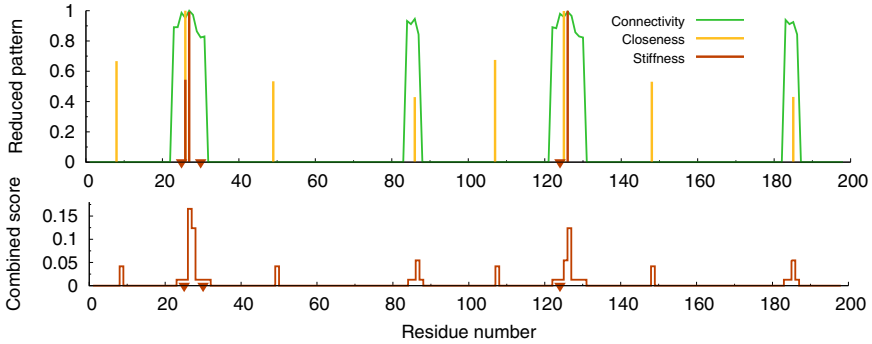

Figure 7. Analysis of HIV-1 protease (PDB 1A30). The upper plot shows the three reduced indicator patterns. The bottom panel illustrates the combined site score given by eq. (9).

where i denotes the amino acid site and σ_i are renormalized indicator patterns, where each peak has the same height $1/N_p^{\mathcal{I}}$ ($N_p^{\mathcal{I}}$ is the number of peaks in the pattern of indicator $\mathcal{I}$), so that $\sum_i \sigma_i = 1$.

By construction, one has $\sum_i S_i = 1$. The meaning of S_i is to gauge the local prediction by counting the number of peaks present in the three indicator patterns, for the sake of simplicity each of them counted with an equal weight of 1/3. Within each profile, each peak is assigned an *internal* weight inversely proportional to the overall number of peaks. Again, the idea is that the larger the number of peaks, the easiest is to make a prediction at some site and consequently the less significative the prediction itself. Furthermore, the site scores S_i can be combined to produce an overall score $S_{\Delta n}$ for a given enzyme by adding up all the scores within Δn sites from the known N_c catalytic sites, that is,

$$S_{\Delta n} = \sum_{i_c=1}^{N_c} \sum_{|i-i_c|=\Delta n} S_i \tag{10}$$

If $S_{\Delta n} > 0$ for a given structure, our algorithm is able to provide at least one prediction. An analysis performed over the whole CSA database shows that the fraction of structures where the combined algorithm returns a prediction is 0.61 for $\Delta n = 1$ and 0.68 for $\Delta n = 2$ (see also Supplementary Material).

In order to elucidate the meaning of the site scores S_i and global score $S_{\Delta n}$, it proves useful to concentrate on a specific enzyme. In Fig. 7, we consider the classic case of HIV-1 protease. Let us first concentrate on the profile of the combined score (9). Two facts are immediately apparent: (i) the catalytic sites appear to be all captured but (ii) there are a number of *orphan* peaks. The global scores for this enzyme are $S_{\Delta n=1} = 0.27$ and $S_{\Delta n=2} = 0.29$. Thus, despite the algorithm flags correctly all the catalytic sites, it does so with some degree of over-prediction (incidentally, we observe that the orphan peaks shown in the combined score profile in Fig. 7 might as well spotlight some hitherto unknown functional sites of HIV-1 protease). Of course, when applying the algorithm to unannotated structures, one does not know *a priori* which peak in the combined score is more likely to point to a catalytic site. This shows the limitations of using *only* a combined score. Analogous conclusions can be drawn by looking at the fraction of catalytic sites predicted by one or more indicators (see supplementary material). For example, closeness and stiffness reduced patterns predict 52% and 28%, respectively, of the catalytic sites within a range $\Delta n = 2$. However only a fraction of 22% is predicted by both. Our conclusion is that each indicator has its specific predictive power, which should be exploited independently, while combined scores should be checked to gauge the *confidence* associated with multiple-indicator predictions.

Looking again at the example of HIV protease will make our point more clear (Fig. 7). It is not difficult to realize that a *sequential* inspection of the three separate indicator profiles at their respective optimal cutoff values is more likely to point to the known catalytic sites first. By inference, we propose that the same inspection sequence be adopted for hitherto unannotated proteins. The connectivity profiles should be examined first. These are the ones with the largest number of peaks, often coalescing to highlight extended regions. The search should be subsequently narrowed down with the corresponding closeness profile, typically featuring more localized peaks, albeit many of them likely to be orphan ones. The prediction should then be refined through the reduced stiffness patterns, the ones with the least number of peaks. Of course, extra information coming from other structure- and/or sequence-based algorithms should be used at each step in conjunction with our algorithm, if possible, to single out interesting sites.

As a final observation, we note that our choice to attribute an equal weight to the three indicators in constructing the combined score S_i in eq. (9) is arbitrary. It would be interesting to inquire whether there exists an optimal combination of weights $w_{\mathcal{I}}$ defining better generalized scores, namely $\mathcal{S}_i = w_\chi \sigma_i^\chi + w_{CC}\sigma_i^{CC} + w_c\sigma_i^c$ with $\sum_{\mathcal{I}} w_{\mathcal{I}} = 1$. For example, one may imagine to use standard optimization techniques[77,78] or genetic algorithms[79] to efficiently determine an optimal set of weights, by training our algorithm on the CSA and other databases.

References

1. Dukka, B. K. C. Structure-based methods for computational protein functional site prediction. *Comp. and Struct. Biotech. J.* **8,** 1–8 (2013).
2. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. prot.* **5,** 725–738 (2010).
3. Laurie, A. & Jackson, R. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21,** 1908–1916 (2005).
4. Neuvirth, H., Raz, R. & Schreiber, G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338,** 181–199 (2004).
5. Puntervoll, P. *et al.* ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucl. Ac. Res.* **31,** 3625–3630 (2003).
6. Devos, D. & Valencia, A. Practical limits of function prediction. *Prot.: Struct., Funct. and Gen.* **41,** 98–107 (2000).
7. Olivier, L., R., B. H. & E., C. F. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *J. Mol. Biol.* **257,** 342–358 (1996).
8. Zhang, T. *et al.* Accurate sequence-based prediction of catalytic residues. *Bioinformatics* **24,** 2329–2338 (2008).
9. Fischer, J. D., Mayer, C. E. & Söding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **24,** 613–620 (2008).
10. Capra, J. A. & Mona, S. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23,** 1875–1882 (2007).
11. Petrova, N. & Wu, C. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics* **7** (2006).
12. Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. *Curr. Op. Struct. Biol.* **15,** 275–284 (2005).
13. Panchenko, A. R., Kondrashov, F. & Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Prot. Sci.* **13,** 884–892 (2004).
14. Innis, C., Anand, A. P. & Sowdhamini, R. Prediction of Functional Sites in Proteins Using Conserved Functional Group Analysis. *J. Mol. Biol.* **337,** 1053–1068 (2004).
15. Sankararaman, S., Sha, F., Jack, F. K., Michael, I. J. & Sjölander, K. Active site prediction using evolutionary and structural information. *Bioinformatics* **26,** 617–624 (2010).
16. Thibert, B., Bredesen, D. & Del Rio, G. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics* **6,** 1–15 (2005).
17. Cheng, G., Qian, B., Samudrala, R. & Baker, D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucl. Ac. Res.* **33,** 5861–5867 (2005).
18. Madabushi, S. *et al.* Structural clusters of evolutionary trace residues are statistically significant and common in proteins1. *J. Mol. Biol.* **316,** 139–154 (2002).
19. Tong, W. *et al.* Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Prot. Sci.* **17,** 333–341 (2008).
20. Ko, J. *et al.* Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. *Prot.: Struct., Funct. Bioinf.* **59,** 183–195 (2005).
21. Gutteridge, A., Bartlett Gail, J. & Thornton Janet, M. Using A Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes. *J. Mol. Biol.* **330,** 719–734 (2003).
22. Riniker, S., Allison, J. R. & Van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: a review. *Phys. Chem. Chem. Phys.* **14,** 12423 (2012).
23. Tozzini, V. Coarse-grained models for proteins. *Curr. Op. Struct. Biol.* **15,** 144–150 (2005).
24. Zheng, W. & Tekpinar, M. Large-scale evaluation of dynamically important residues in proteins predicted by the perturbation analysis of a coarse-grained elastic model. *BMC Struct. Biol.* **9** (2009).
25. Demerdash, O. N. A., Daily, M. D. & Mitchell, J. C. Structure-Based Predictive Models for Allosteric Hot Spots. *PLoS Comp. Biol.* **5** (2009).
26. Haliloglu, T., Seyrek, E. & Erman, B. Prediction of Binding Sites in Receptor-Ligand Complexes with the Gaussian Network Model. *Phys. Rev. Lett.* **100,** 228102 (2008).
27. Emekli, U., Schneidman-Duhovny, D., Wolfson, H. J., Nussinov, R. & Haliloglu, T. HingeProt: Automated prediction of hinges in protein structures. *Prot.: Struct., Funct. and Bioinf.* **70,** 1219–1227 (2008).
28. Ertekin, A., Nussinov, R. & Haliloglu, T. Association of putative concave protein-binding sites with the fluctuation behavior of residues. *Prot. Sci.* **15,** 2265–2277 (2006).
29. Yang, L. W. & Bahar, I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* **13,** 893–904 (2005).
30. Tirion, M. M. Low-amplitude elastic motions in proteins from a single-parameter atomic analysis. *Phys. Rev. Lett.* **77,** 1905–1908 (1996).
31. Haliloglu, T., Bahar, I. & Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **79** (1997).
32. Atilgan, A. *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80,** 505–515 (2001).
33. Piazza, F., Rios, P. D. L. & Cecconi, F. Temperature dependence of normal mode reconstructions of protein dynamics. *Phys. Rev. Lett.* **102,** 218104–4 (2009).
34. Bahar, I. & Cui, Q. (eds.) *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. C&H/CRC Mathematical & Computational Biology Series*. vol. 9 (CRC press, Boca Raton, 2005).
35. De Los Rios, P. *et al.* Functional dynamics of pdz binding domains: A normal-mode analysis. *Biophys. J.* **89,** 14–21 (2005).
36. Delarue, M. & Sanejouand, Y.-H. Simplified normal modes analysis of conformational transitions in dna-dependant polymerases: the elastic network model. *J. Mol. Biol.* **320,** 1011–1024 (2002).
37. Tama, F. & Sanejouand, Y.-H. Conformational change of proteins arising from normal mode calculations. *Prot. Eng.* **14,** 1–6 (2001).
38. Piazza, F. & Sanejouand, Y.-H. Energy transfer in nonlinear network models of proteins. *Europhys. Lett.* **88,** 68001 (2009).
39. Piazza, F. & Sanejouand, Y.-H. Long-range energy transfer in proteins. *Phys. Biol.* **6,** 046014 (2009).
40. Nicolay, S. & Sanejouand, Y. H. Functional modes of proteins are among the most robust. *Phys. Rev. Lett.* **96** (2006).
41. Bahar, I., Atilgan, A. R., Demirel, M. C. & Erman, B. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.* **80** (1998).
42. Ghoraie, L. S., Burkowski, F. & Zhu, M. Sparse networks of directly coupled, polymorphic, and functional side chains in allosteric proteins. *Prot.: Struct., Funct. and Bioinf.* **83,** 497–516 (2015).
43. Flores, D. I., Sotelo-Mundo, R. R. & Brizuela, C. A. A Simple Extension to the CMASA Method for the Prediction of Catalytic Residues in the Presence of Single Point Mutations. *PloS ONE* **9** (2014).

44. Sanjaka, B. V. M. V. & Yan, C. Prediction of Enzyme Catalytic Sites on Protein Using a Graph Kernel Method. In Chen, L, Zhang, X. S., Wu, L. Y. & Wang, Y. (ed.) 2013 *7th International Conference on Systems Biology (ISB),* IEEE International Conference on Systems Biology, 31–33 (Hunan Univ; ORSC, Computat Syst Biol Soc; IEEE, Syst Biol Tech Comm; IET; Chinese Acad Sci; Natl Nat Sci Fdn China; CAS, Acad Math & Syst Sci; CAS, Shanghai Inst Biol Sci, 2013). 7th International Conference on Systems Biology (ISB), Huangshan, PEOPLES R CHINA, AUG 23–25 (2013).
45. Fajardo, E. J. & Fiser, A. Protein structure based prediction of catalytic residues. *BMC Bioinformatics* **14,** 63 (2013).
46. Gonzalez, A. J., Liao, L. & Wu, C. H. Predicting Ligand Binding Residues and Functional Sites Using Multipositional Correlations with Graph Theoretic Clustering and Kernel CCA. *IEEE-ACM Trans. on Comp. Biol. and Bioinf.* **9,** 992–1001 (2012).
47. Pons, C., Glaser, F. & Fernandez-Recio, J. Prediction of protein-binding areas by small-world residue networks and application to docking. *BMC Bioinformatics* **12,** 378 (2011).
48. Vacic, V., Iakoucheva, L. M., Lonardi, S. & Radivojac, P. Graphlet Kernels for Prediction of Functional Residues in Protein Structures. *J. Comp. Biol.* **17,** 55–72 (2010).
49. Li, G.-H. & Huang, J.-F. CMASA: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation. *BMC bioinformatics* **11,** 439 (2010).
50. Amitai, G. *et al.* Network Analysis of Protein Structures Identifies Functional Residues. *J. Mol. Biol.* **344,** 1135–1146 (2004).
51. Sharp, K. & Skinner, J. J. Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Prot.: Struct., Funct. and Bioinf.* **65,** 347–361 (2006).
52. Tang, Y.-R., Sheng, Z.-Y., Chen, Y.-Z. & Zhang, Z. An improved prediction of catalytic residues in enzyme structures. *Prot. Eng. Des. Select.* **21,** 295–302 (2008).
53. Slama, P., Filippis, I. & Lappe, M. Detection of protein catalytic residues at high precision using local network properties. *BMC Bioinformatics* **9** (2008).
54. Chea, E. & Livesay, D. R. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics* **8,** 153 (2007).
55. Zvelebil Markéta, J. J. M. & Sternberg, M. J. Analysis and prediction of the location of catalytic residues in enzymes. *Prot. Eng.* **2,** 127–138 (1988).
56. Bate, P. & Warwicker, J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.* **340,** 263–76 (2004).
57. Elcock, A. H. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **312,** 885–96 (2001).
58. Sacquin-Mora, S., Delalande, O. & Baaden, M. Functional modes and residue flexibility control the anisotropic response of guanylate kinase to mechanical stress. *Biophys. J.* **99,** 3412–3419 (2010).
59. Sacquin-Mora, S., Laforet, E. & Lavery, R. Locating the active sites of enzymes using mechanical properties. *Proteins* **67,** 350–359 (2007).
60. Brodkin, H. R. *et al.* Prediction of distal residue participation in enzyme catalysis. *Prot. Sci.* **24,** 762–778 (2015).
61. Lee, J. & Goodey, N. M. Catalytic contributions from remote regions of enzyme structure. *Chem. Rev.* **111,** 7595–7624 (2011).
62. Piazza, F. & Sanejouand, Y.-H. Discrete breathers in protein structures. *Phys. Biol.* **5,** 026001 (2008).
63. Juanico, B., Sanejouand, Y.-H., Piazza, F. & De Los Rios, P. Discrete breathers in nonlinear network models of proteins. *Phys. Rev. Lett.* **99,** 238104 (2007).
64. Kondrashov, D. A., Cui, Q. & Phillips, G. N. J. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophys. J.* **91,** 2760–2767 (2006).
65. Suhre, K. & Sanejouand, Y.-H. Elnémo: a normal mode server for protein movement analysis and the generation of templates for molecular replacement. *Nucl. Ac. Res.* **32,** W610–W614 (2004).
66. Hafner, J. & Zheng, W. Optimal modeling of atomic fluctuations in protein crystal structures for weak crystal contact interactions. *J. Chem. Phys.* **132,** 014111 (2010).
67. Riccardi, D., Cui, Q. & Phillips, J. George N. Application of elastic network models to proteins in the crystalline state. *Biophys. J.* **96,** 464–475 (2009).
68. Eyal, E., Yang, L.-W. & Bahar, I. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **22,** 2619–2627 (2006).
69. Rader, J. A. & Brown, M. S. Correlating allostery with rigidity. *Mol. BioSys.* **7,** 464–471 (2011).
70. Kamal, M. Z., Mohammad, T. A. S., Krishnamoorthy, G. & Rao, N. M. Role of Active Site Rigidity in Activity: MD Simulation and Fluorescence Study on a Lipase Mutant. *PLoS ONE* **7,** e35188 (2012).
71. Guo, X. *et al.* Strain energy in enzyme-substrate binding: An energetic insight into the flexibility versus rigidity of enzyme active site. *Comp. Theo. Chem.* **995,** 17–23 (2012).
72. Brandman, R., Lampe, J. N., Brandman, Y. & De Montellano, P. R. O. Active-site residues move independently from the rest of the protein in a 200 ns molecular dynamics simulation of cytochrome p450 cyp119. *Arch. Biochem. Biophys.* **509,** 127–132 (2011).
73. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35,** 539–46 (2010).
74. Bertil, H. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA* **99,** 1274–1279 (2002).
75. Porter, C. T., Bartlett, G. J. & Thornton, J. M. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl. Ac. Res.* **32,** D129–D133 (2004).
76. Lanczos, C. *Applied Analysis* (Prentice-Hall, Englewood Cliffs, New England, 1956).
77. Diwekar, U. *Introduction to Applied Optimization* (Kluwer Academic Publishers, 2003).
78. Bertsimas, D. & Tsitsiklis, J. *Introduction to Linear Optimization* (Athena Scientific, Belmont, MA, 1997).
79. Reeves, C. & Rowe, J. *Genetic Algorithms: Principles and Perspectives* (Kluwer Academic Publishers, 2002).

Author Contributions

F.P. conceived the study. S.A. performed the calculations. All authors analyzed the results and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Aubailly, S. and Piazza, F. Cutoff lensing: predicting catalytic sites in enzymes. *Sci. Rep.* **5**, 14874; doi: 10.1038/srep14874 (2015).