Using A Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes

Alex Gutteridge, Gail J. Bartlett and Janet M. Thornton*

EBI, Wellcome Trust Genome Campus, EMBL Outstation
Hinxton, Cambridgeshire CB10 1SD, UK

Structural genomics projects aim to provide a sharp increase in the number of structures of functionally unannotated, and largely unstudied, proteins. Algorithms and tools capable of deriving information about the nature, and location, of functional sites within a structure are increasingly useful therefore. Here, a neural network is trained to identify the catalytic residues found in enzymes, based on an analysis of the structure and sequence. The neural network output, and spatial clustering of the highly scoring residues are then used to predict the location of the active site.

A comparison of the performance of differently trained neural networks is presented that shows how information from sequence and structure come together to improve the prediction accuracy of the network. Spatial clustering of the network results provides a reliable way of finding likely active sites. In over 69% of the test cases the active site is correctly predicted, and a further 25% are partially correctly predicted. The failures are generally due to the poor quality of the automatically generated sequence alignments.

We also present predictions identifying the active site, and potential functional residues in five recently solved enzyme structures, not used in developing the method. The method correctly identifies the putative active site in each case. In most cases the likely functional residues are identified correctly, as well as some potentially novel functional groups.

© 2003 Elsevier Science Ltd. All rights reserved

Keywords: bioinformatics; structural genomics; functional prediction; neural network; active sites

Introduction

The huge increase in the rate of DNA sequencing, and the use of gene prediction technologies, such as Genscan and GeneWise, have flooded protein databases with new sequence data. The various structural genomics initiatives (SGI), now aim to produce a similar increase in the amount of structural information. One of the most important tasks in biology today is to use these data to provide functional annotation that leads to biologically useful knowledge.

One type of information that structural data can provide is the location and nature of the functional regions of a protein, such as protein–protein interaction sites and ligand binding pockets. Knowing the location of the functional sites within a protein allows the study of targeted mutants, structure-based drug design, and functional annotation of the protein by comparison with other characterised proteins.

Most novel genes are functionally annotated using sequence analysis to find similar genes of known function, typically by running one of the various flavours of BLAST or profile and HMM approaches such as Pfam. Several studies have pointed out the problems associated with homology based functional annotation and indicate a cut-off of sequence identity as high as 40%, below which it is dangerous to transfer anything but the
broadest functional annotation. It is certain that some sequences in the public databases are incorrectly annotated due to the difficulty of transferring function based purely on sequence similarity.

As more data comes through from structural genomics it is likely that a similar approach will be taken to annotate proteins using structural homologues. The idea that proteins with similar structures perform similar functions has been examined closely. Generally, although it is true that structure is more conserved than sequence at great evolutionary distance, the transfer of function based on structural similarity is no more reliable than annotation based on sequence similarity. The problem being the limited number of unique folds found in nature, which has been estimated to be as low as 1000. Since the number of functions performed by proteins far exceeds this number, it follows that one fold must be capable of many functions. The triose phosphate isomerase (TIM) barrel fold, for instance, is associated with 61 different EC numbers, covering five of the six top level EC classifications.

Methods to locate and characterise the functional sites of a protein could provide data for functional annotation in ways not based on homology, as well as providing information for mutagenesis and drug design studies. Traditional molecular biology techniques for finding functional sites, such as mutagenesis, are generally time consuming, and rely on some prior knowledge of the function of the protein to allow it to be assayed.

In silico methods for finding and annotating functional sites would clearly be of great help in annotating novel protein structures from structural genomics. Several different strategies have already been developed, however, none has been used to perform an analysis across the whole structure database. Pattern matching approaches such as TESS, FFF and SPASM aim to locate functional sites and annotate structures by finding small three-dimensional motifs within the structure. The disadvantage of this is that suitable motifs have to be derived (usually from literature, though data-mining techniques have been used for automatic extraction of motifs) and truly novel structures may not match any known motif. Recent studies have also presented methods for finding similarities between cavities on the protein surface which could be used to annotate structures, once the functionally important cavities are identified.

Techniques for finding functional sites de novo, such as evolutionary trace (ET), and other similar methods generally focus on searching for three-dimensional clusters of conserved residues. ET studies have made genuine, experimentally confirmed, predictions for the location of functional sites in G-proteins and regulator of G-protein signalling (RGS) proteins demonstrating the potential of this type of technique.

Some proteins, including those targeted by structural genomics, have no sequence homologues and so conservation based approaches do not work. Methods, which only use structural information to locate functional residues, have been developed to provide functional annotation for these proteins. These techniques rely on identifying residues with unusual electrostatic and ionisation properties, and have shown a correlation between these residues and functional sites within the protein.

Here, we describe a new method for de novo prediction of functional sites specific for the active sites of enzymes. Instead of searching for clusters of conserved residues, a neural network is used to score the residues of a protein structure by the likelihood that they are catalytic. By searching for clusters of high-ranking residues the algorithm determines the most likely active site. The neural network is trained using a dataset of proteins for which the catalytic residues have been confidently located by experiment. Structural parameters such as the solvent accessibility, type of secondary structure, depth, and cleft that the residue lies in, as well as the conservation score and residue type are used as inputs for the neural network.

Results
Analysis of parameters

A detailed analysis of the parameters is provided by Bartlett et al. A brief summary is presented here.

Conservation was the most powerful parameter for discriminating catalytic and non-catalytic residues. Some proteins, however, failed to find sufficiently diverse homologues to generate meaningful conservation scores. It is hoped that in these cases the predictive power of the other parameters will be enough to allow reliable predictions to still be made.

Catalytic residues show a tendency to be buried within the structure and so have a lower relative solvent accessibility (RSA) than other residues, particularly non-catalytic polar residues, the majority of which lie exposed on the surface of the protein. Despite this tendency to be buried, catalytic residues are often found lining a large cleft. This tendency is particularly marked for the largest cleft, and is significant for the second and third largest clefts. For clefts smaller than this (fourth to ninth largest) the difference is not particularly significant.

There is a slight tendency for catalytic residues to prefer coil regions over helix or sheet regions, this could be due to the extra conformational flexibility this gives them (allowing the active site to change conformation on ligand binding).

The hydrophobic and small residue groups were found to be very rarely catalytic, presumably because they do not contain the chemical groups required for most catalytic tasks. The obvious
exceptions to this are when the backbone amide and carbonyl groups perform catalytic functions. It was found that glycine is the residue most often used in this case.

Depth

Depth values were calculated for the non-catalytic residues in the data set, and the distribution (Figure 1 shows that almost 40% of residues lie on, or near, the surface of the protein, with depths less than 1 Å. These residues are almost completely exposed to the surface with only a few of their atoms not solvent accessible. The proportion of the total represented by each 1 Å division then decreases steadily, apart from a small-peak in the 4–5 Å division. Presumably this second-peak is due to invaginations on the protein surface, which alter the distribution from the smooth decrease one would expect given a perfectly spherical protein. The very deepest residues in this data set lie at ∼13 Å. Catalytic residues show a different distribution, with only 17% lying in the outer 1 Å, the majority occupy the next partially buried layer between 2 Å and 4 Å. This allows the catalytic residues to have some solvent accessibility (in order to interact with the substrates) whilst remaining mostly buried (to allow themselves to be correctly orientated by other residues). The catalytic residues rarely have depths greater than 5 Å.

An example: quinolinate phosphoribosyltransferase

As an example of the neural network output, the scores along the 286 amino acid sequence of quinolinate phosphoribosyltransferase (1QPR) are shown in Figure 2. Most residues score very low (a large majority score less than 0.01), and around 20 residues score over 0.5. The four known catalytic residues (Arg105, Lys140, Glu201 and Asp222) all score highly, though several other residues score as high or higher. There is some grouping of the high-scoring residues in the sequence, particularly around residue 140, but most high-scores are isolated spikes. When the scores are mapped on to the 1QPR structure (Figure 3) the high-scoring
areas, although widely separated in the sequence, are brought together and cluster into two areas corresponding to the two active sites of the quinolate phosphoribosyltransferase homodimer.

Figure 3. (a) Distribution of neural network scores in the 1QPR structure. Residues are coloured by network score (Red = high, blue = low). (b) The structure of the 1QPR homodimer, coloured by chain, with the known catalytic residues drawn in thick lines. All structure diagrams are prepared using PyMol.

Training the network

The training process is tracked by measuring the Matthews correlation coefficient (MCC) after each epoch. Figures 4 and 5 show how the MCC varies as training progresses. The variation in performance is quite considerable, with the final MCC varying between 0.35 and 0.25, reflecting the natural variation within the data set. Figure 5 shows the MCC varying with each epoch averaged over all ten runs. The network reaches its best MCC after only 30 epochs or so, levelling off at an average MCC of around 0.28. There is no evidence of over-fitting in the results, as the MCC does not fall significantly once it has plateaued.

Network weights

The relative strength of the weights that the network converges to are shown in Figure 6. Conservation and diversity of position score (DOPS) are both highly weighted. As expected the network also looks for buried residues, as RSA is given a negative weighting. The cleft categories show that lying in a cleft, and the size of that cleft are important factors in the network score, though not important as conservation or RSA. Depth is not weighted strongly in either direction, and is not important in making a prediction. The difference for the secondary structure parameters is also small. Residue type has a very large variation with histidine, cysteine and the charged residues (aspartate, glutamate, lysine and arginine residues) all scoring highly, whilst the hydrophobic residues score low.

The high DOPS weighting is interesting, as it is the same for all residues within a protein chain. The only effect is to raise all the scores of all residues in chains with high DOPS and lower all the scores of all residues in chains with low DOPS. The network has learnt that when DOPS is low it is better, in terms of the overall error rate, to make no catalytic predictions at all, rather than predict everything to be catalytic. Since the clustering algorithm uses residues based on their rank rather than absolute scores, this makes no difference in the later stages.

Figure 4. Training the neural network, each line represents one of the ten cross validation runs.
Clustering

In the network scoring we consider each residue as independent of the others, however, catalytic residues are likely to cluster together in the structure. Ranking and clustering the residues allows us to use this information to improve the predictions and locate the active site. For each structure a list of possible catalytic residues is generated by ranking the residues by network score. The clustering algorithm finds distinct clusters of these residues and generates a sphere that forms the predicted active site. 1158 clusters are generated from the test set, an average of 7.2 per protein. The multimeric nature of most of the proteins means that the average number of known active sites is 2.6 per protein. The distribution of sphere sizes for the known sites and all the predicted sites is shown in Figure 7. Figure 8 shows the sizes for the known sites and the top scoring predicted sites only.

Most predicted clusters are small and contain two or three members with a radius of 3–4 Å, in contrast the top scoring predictions in each structure are generally large and lie at the upper end of the allowed size range (15 Å). The known sites generate spheres with sizes between 6 Å and 12 Å, though a significant number have a single catalytic residue and so have radii of 3 Å. A few outliers have spheres larger than 20 Å in radius. These cases all represent structures where the catalytic cluster is thought to come together upon substrate binding so the cluster appears very large in the unbound form.

Comparing the predicted sites to the known sites

To test whether a prediction is correct, the overlap between the predicted site and the closest known active site is calculated. A correct prediction occurs

![Figure 5. MCC averaged over all ten cross validation runs.](image)

![Figure 6. The relative strengths of the weights placed on the various parameters. Categorical parameters such as residue type are grouped, with the lowest weight set at 0.](image)
when the overlap is greater than 50% of the volume of the known active site, a partially correct prediction occurs when there is some overlap but less than 50%, a failure occurs when there is no overlap between the known and predicted spheres.

For each protein in the test set, the prediction with the highest total network score was selected and compared to the known sites. The results are shown in Figure 9, 62% of the proteins have the active site correctly identified, and a further 22% are partially correct.

When we consider the overlap for all the sites predicted for each protein we find the results improve: 69% of the proteins have the active site correctly identified and 25% have a partially correct prediction. The increase of only 7% when all predictions are considered shows that the highest scoring cluster is very often the true active site. Eleven cases were found where the top prediction was not correct, but one of the other predictions was. In six of these cases the correct cluster was the second or third highest scoring prediction and in four cases the correct cluster was the fourth or fifth highest scoring, in the final case the correct cluster was the seventh highest scoring.

When each of the 1158 predicted clusters is considered individually, as opposed to by each protein, 25% are found to be correct and 41%
are partially correct. The high number of partial hits is presumably due to the tendency of the network to find residues lying near the active site, but which aren’t close enough to the true catalytic residues to score as correct. It is also possible that many of the partially correct and incorrect clusters represent secondary functional sites such as ligand binding or protein–protein interaction sites. These clusters are biologically interesting, but are considered incorrect when searching solely for active sites.

**Significance of results**

To calculate the significance of these results, we estimate the probability \( P_R \) of achieving this level of prediction by random chance. A similar method to that used by Aloy et al.\(^{26} \) is applied. To a reasonable approximation a correct hit occurs when the centre of the smaller of the two spheres lies within the volume of the larger. Since the known catalytic site is usually smaller than the largest predicted site, and assuming the prediction has an equal probability of being anywhere within the volume of the protein, \( P_R \) is the ratio of the volume of the predicted sphere to the volume of the protein. Since most of the proteins are multimeric this ratio is then multiplied by the number of active sites (any one of which could have overlapped with the predicted site).

The volume of each protein is estimated by drawing a sphere around all the C\(^\text{\textbf{b}}\) atoms of the structure, giving an average of 510,000 Å\(^3\). Since most catalytic residues lie in the outer 5 Å of the protein we shall consider the predictions restricted to only a third of this volume. The average volume of all the predicted spheres is 2632 Å\(^3\), and the average volume of the top scoring predictions is 5783 Å\(^3\). There are 7.2 predicted sites and 2.6 known sites per protein on average. A summary of the observed and expected rates of correct predictions for the three different analysis is shown in Table 1. We estimate the significance of the differences using equation (1),\(^{26} \) which follows a normal distribution with mean 0 and standard deviation 1. All the results are significant to more than 10\(^{-15} \).

\[
z = \frac{P_O - P_R}{\sqrt{\frac{P_R(1 - P_R)}{n}}} \quad (1)
\]

**Comparison of the performance of different networks**

The neural network and clustering process incorporates a variety of different types of information: evolutionary information encoded in the conservation scores, residue propensities, structural information in the parameters and detailed structural information included in the clustering stage. To understand how these different types of information contribute to the overall performance, networks have been trained using different subsets of the parameters.

Two additional networks have been developed. First a network trained solely using sequence parameters (conservation, DOPS and residue propensities), and second a network trained using structural parameters but excluding conservation and DOPS scores. Residue propensity is included in the structural information as the sequence of a protein would always be known given a structure. The relative performance of the different networks is shown in Figure 10 and in detail in Table 2. The performance of each network in finding the location of the active site is shown in Figure 11 and Table 3.

The performance of the technique described by Aloy et al.,\(^{26} \) which uses conservation, residue propensity and clustering is also shown in Table 3. This study uses the same sphere based method as shown here to assess the accuracy of the predictions making comparison easy. The functional residues were based on SITE records in PDB files, which are not as well defined as the catalytic residues used here. Aloy et al. analysed 106 proteins and found that 20 of them could not generate sufficiently diverse alignments to give good predictions. Since, here, we have included proteins with low DOPS, we include these 20 proteins as incorrect predictions when calculating performance. Once this is taken into account we find that of the 106 proteins, 68 are correctly predicted (64%), 13 partially correct (12%) and 25 are incorrect (24%), when all predictions are considered. This level of prediction is almost identical to the sequence trained network.

**Predictions**

The neural network was run on several recently published enzyme structures, which were not included in the original data, or subsequent analysis,
Figure 10. Comparison of the MCC achieved by the three different networks in predicting catalytic residues, before and after structural clustering is applied.

Figure 11. Comparison of the site prediction accuracy for the three different networks. Results are presented considering all predicted sites and the top scoring site only.

Table 2. Comparison of the performance of the three different neural networks in predicting catalytic residues

<table>
<thead>
<tr>
<th>Data used</th>
<th>Before clustering</th>
<th></th>
<th></th>
<th>After clustering</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCC</td>
<td>Q(predicted)</td>
<td>Q(Observer)</td>
<td>MCC</td>
<td>Q(predicted)</td>
<td>Q(Observer)</td>
</tr>
<tr>
<td>Structure</td>
<td>0.19</td>
<td>0.10</td>
<td>0.41</td>
<td>0.23</td>
<td>0.10</td>
<td>0.57</td>
</tr>
<tr>
<td>Sequence</td>
<td>0.24</td>
<td>0.13</td>
<td>0.50</td>
<td>0.26</td>
<td>0.13</td>
<td>0.58</td>
</tr>
<tr>
<td>Sequence + structure</td>
<td>0.28</td>
<td>0.14</td>
<td>0.56</td>
<td>0.32</td>
<td>0.16</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 3. Comparison of the performance of the three different neural networks in locating active sites

<table>
<thead>
<tr>
<th>Data used</th>
<th>Top sites only (%)</th>
<th></th>
<th></th>
<th>All sites (%)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Partial</td>
<td>Incorrect</td>
<td>Correct</td>
<td>Partial</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Structure</td>
<td>52.8</td>
<td>25.8</td>
<td>19.4</td>
<td>62.3</td>
<td>31.4</td>
<td>8.3</td>
</tr>
<tr>
<td>Sequence</td>
<td>57.2</td>
<td>27.7</td>
<td>15.1</td>
<td>63.5</td>
<td>28.3</td>
<td>8.2</td>
</tr>
<tr>
<td>Sequence + Structure</td>
<td>62.3</td>
<td>21.4</td>
<td>16.4</td>
<td>69.2</td>
<td>24.5</td>
<td>6.3</td>
</tr>
<tr>
<td>Aloy et al.</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>64.2</td>
<td>12.3</td>
<td>23.5</td>
</tr>
</tbody>
</table>
to gauge the usefulness of the method in annotating structures.

**SET domain histone lysine methyltransferases**

Several recent papers have presented the first structures of histone lysine methyltransferase (HMTase) containing SET domains. SET domains are responsible for the methylation of specific lysine residues in histone proteins, leading to changes in chromatin regulation and gene expression. SET domains share no homology with other structurally characterised methyltransferases, and so the structure, and the functional information that the structure contains is of significant importance.

The structure of yeast protein Clr4 (PDB code 1MVX) was used for the prediction of the functional sites. The PSI-BLAST search required 7 iterations to converge (using an E-value cut-off of $10^{-20}$) and found ~150 homologues, producing a very diverse alignment. 31 residues score over the ranking cut-off and clustering reveals one large cluster, containing 19 of these residues. The output of the neural network mapped to the surface of the structure is shown in Figure 12. The dominant cluster forms the large L-shaped patch in the centre of the structure comprising residues His410, Asp450, Asn409, and Tyr451, the other residues in the cluster extend either side of the L-shape patch and into the structure.

Mutations to His410, Cys412, Arg320, Glu446 and Arg406 have been shown to inactivate the enzyme, though it is suggested that Arg320 is most likely to be of structural, rather than catalytic importance. The structure with an AdoHcy cofactor bound is known for homologous SET domains. This reveals that Tyr451, Asn409 and His410 make contacts to the cofactor and Tyr357 is proposed as a possible catalytic proton source. A high resolution crystal structure of the human SET7/9 domain has been recently published. This study suggests catalytic roles for the residues equivalent to Tyr451, Tyr419, Tyr357 and the main-chain carbonyl oxygens of Asp403 and Phe408. Of these functional residues, the large predicted cluster contains His410, Glu446, Arg406, Tyr357, Tyr451 and Asp403. The neural network identifies the correct active site and many of the known functional residues.

**Intron endonuclease I-TevI**

The structure of the intron endonuclease I-TevI from bacteriophage T4 has recently been published. Intron endonucleases catalyse a break in double stranded DNA, that facilitates the insertion of introns and inteins. I-TevI contains separate catalytic and DNA binding domains, the structure of the catalytic domain is analysed here (PDB code: 1LN0). Using the default PSI-BLAST parameters the sequence of 1LN0 picks up no homologues.

Figure 12. (a) The front face of the SET domain showing the large, high-scoring surface patch and the Ado-HCys binding cleft. Residues His410, Asp450, Asn409 and Tyr451 form the L-shaped patch in the centre, Arg406 and Tyr357 lie in the pocket to the right. (b) The catalytic site of I-TevI, network scores are generated without conservation data. (c) The catalytic site of I-TevI, showing the improved prediction once conservation data is included. (d) l-Arabinanase, the central red patch is made of His37 and Asp38. Asp158 and Glu221 both lie close by in the same pocket. (e) FemA, the binding cleft is the long green patch in the centre of the figure. The most likely catalytic site lies at the far left end of the cleft. (f) The RlmB dimer. The subunits are stacked on top of each other running left to right. The two active sites lie in the high-scoring regions in the interface between the two subunits.
Despite this the network still makes predictions based purely on the structure and the residue propensities. The network identifies three residues (His31, His40, Ser42) forming the highest scoring cluster.

A putative active site is proposed based on conservation and mutagenesis data. The site is located in the same cleft identified by the network. Glu75 binds a divalent cation and is likely to be the principal functional residue. Other functional residues suggested by the authors include Tyr17, Arg27, His31 and His40. The 1LN0 structure has Arg27 mutated to alanine, as active I-TevI cannot be produced by *Escherichia coli*. Replacing Ala27 by arginine in the sequence presented to PSI-BLAST, and reducing the E-value cut-off to $10^{-5}$, allows the network to improve the prediction. Twelve residues now form the largest cluster including Tyr17, Arg27, His31, and His40. Glu75 still remains outside the predicted cluster, however.

This example demonstrates how the network can cope with structures occupying a sparsely populated region of sequence space. The prediction made only on the basis of residue propensities and structural data correctly identifies the active site and several functional residues. Once the mutated structure is corrected and conservation scores are added the network makes improved predictions, correctly identifying the active site and many of the principal residues, though it still fails to predict the crucial Glu75. The problems of mutated structures and limited sequence homologues highlight some of the difficulties that would be encountered in a PDB-wide analysis.

**α-L-Arabinanase**

The structure of *Cellvibrio japonicus* arabinanase has been solved recently revealing a novel five-bladed β-propeller fold (PDB code: 1GYD). Arabinanase hydrolyses the arabinans polymers found in plant cell walls. The PSI-BLAST search converges after only four iterations, only finding 11 homologues and generating a very diverse alignment. The network scores mapped to the structure are shown in Figure 12. The high-scoring residues line the large cleft that runs the length of the protein. The clustering algorithm suggests a seven residue cluster comprising the high-scoring residues His106 and His29, and five other lower scoring residues. This cluster lies at the very end of the cleft. Another five residue cluster lies approximately halfway along the cleft comprising Lys383, Phe382, Ser342, Ser314 and Thr352. The crystal structure does not have any ligand bound, and no mutagenesis data is available to pinpoint the actual catalytic residues.

The cleft is the only structure large enough to accommodate the peptidoglycan substrate and hence is the most likely binding site, though a conformational change on substrate binding cannot be ruled out. The network suggests several residues as potential catalytic groups and further experimentation is required to confirm which, if any, of these residues form the catalytic centre.

**RlmB 23 S rRNA Methyltransferase**

RlmB is an *Escherichia coli* protein representing the novel Ado-Met dependent methyltransferase class, SPOUT. RlmB is responsible for the methylation of a specific guanosine group in the 23 S rRNA component of the ribosome. The crystal structure of the enzyme has recently been solved (PDB code: 1GZ0). PSI-BLAST converges after iteration five, having found 100 homologues and generates a very diverse alignment. RlmB forms a homodimer in solution and the high-scoring residues cluster into two almost identical sites in the dimer interface region. Each site contains residues from both chains A and F. The highest scoring residue is Arg114 which is involved in a salt-bridge with Glu198 from the opposite chain. Surrounding this pair are His9, Asp117, Glu147, Ser148 and Gly144 from the same chain as Arg114 and Ser224, Leu225, Asn226 and Ser228 from the same chain as Glu198. A secondary cluster comprised of Asp105, His291, Ser292 and Thr353 lies 4.3 Å from this main cluster.

The authors propose a putative active site based on conservation of three previously identified motifs, found in most methyltransferases. Motif 1 covers residues Asn108 to Arg114, motif II...
covers Glu198 and motif III covers Ser224, Leu225 and Asn226. They also report that mutagenesis of the equivalent residue to Glu198 in a homologue abolishes methyltransferase activity. Glu198 and Ser224 are suggested as possible catalytic bases. His9 is implicated in RNA binding, however, several other putative RNA binding residues are not identified strongly by the network.

The network has correctly identified the putative catalytic centre, though again the clustering has split the site, leaving part in a small secondary cluster.

Discussion

One of the original aims of the project, to predict catalytic residues from structures, has proven to be an extremely difficult task given the narrow definition of catalytic used here. The MCC of 0.28 (or 0.32 if clustering is used) is too low to realistically use the simple predictions from the neural network in identifying catalytic residues directly. The main problem is the high number of false positives. 56% of catalytic residues are identified correctly, but only one in seven catalytic predictions are correct.

Visual inspection of the results shows that many of the false positives are other functional residues lying in the active site such as substrate binding and metal binding residues. These residues have very similar properties to the catalytic residues: conserved, low-solvent accessibility, lying in clefts and they also lie extremely close to the true catalytic residues and do not form a distinct or separate spatial cluster. A system looking to identify any functional residues at the active site may well consider these false positives to be true positives, however, given the definition used here they are errors. As well as the problem of these false positives there is the inherent difficulty of picking the handful of catalytic residues from hundreds in the protein. The ratio of catalytic to non-catalytic is around one in one hundred across the entire data set. Given these difficulties the low success rate is understandable and not as disappointing as first appears.

The network weights and the performance of the sequence-only neural network shows that evolutionary information, encoded in conservation scores is very important in making a prediction. This network reflects the performance that one could expect to achieve when predicting catalytic residues purely from sequence data. We see from the $Q_{\text{observed}}$ and $Q_{\text{predicted}}$ values in Table 2 that 50% of catalytic residues are found by this network, but only one in eight of the predictions is correct.

Structural genomics projects aim to provide some level of structural information for the majority of protein sequences. Some of these proteins will not have any known sequence homologues and the structure will be the only information available. The neural network trained without conservation scores reflects the performance one could expect to achieve when analysing these proteins. The network alone performs poorly, however, the structural information can also be used to cluster the predictions in these proteins. When this form of structural information is incorporated the overall performance rises almost to the level of the sequence network, and 57% of the catalytic residues are correctly predicted, though the true positives are still only one in ten of the catalytic predictions.

For the majority of structural genomics targets there are some sequence homologues and in these cases both types of information can be incorporated. The network trained using sequence and structure outperforms both the other networks with an MCC of 0.28 rising to 0.32 when clustering is used (Table 2). 68% of catalytic residues are correctly predicted and one in six of the catalytic predictions is correct.

Although predicting the catalytic residues is difficult, predicting the location of the active site can be done with significant levels of success (Table 3). When only structural information is used the clustering algorithm is still able to correctly identify the catalytic cluster in 62% of proteins and a partially correctly in a further 31%. This suggests that even for structural genomics targets where no conservation data is available, it will still be possible to make significant predictions about the location of the active site.

The neural network trained using sequence data identifies 63.5% of sites when all predictions are considered. This level of performance is similar to the technique described by Aloy et al. and which also uses conservation, residue propensities and clustering. It should be noted that Aloy et al. compared their predictions to the SITE records of PDB files, which are less rigorously defined than the catalytic clusters used here, and generally comprise larger number of residues. The performance of the neural networks used here are likely to be underestimated compared to Aloy et al., therefore.

As with the neural network output, when structure and sequence are combined the performance exceeds that of sequence or structure alone. In this case 69% of sites are correct considering all predictions and 62% considering only the top prediction. A further 25% of sites are partially correctly predicted when all predictions are considered and 22% when only the top prediction is considered. The method fails to make a useful prediction in only 6% of cases when all the predictions are examined.

One of the justifications for the large investment made in structural genomics is that it will allow identification of functional sites and residues in cases where it is not possible from sequence. The results we have shown here indicate that structure alone can be used to identify catalytic residues and active sites in enzymes, however, evolutionary...
history encoded in the form of conservation scores is an extremely rich source of information for making these types of predictions and should be incorporated at every opportunity. The improvement in performance when structure and sequence are used, shows that structural information, other than that used for clustering, should be incorporated into de novo prediction techniques such as evolutionary trace.

Why did the failures fail?

When considering the top scoring sites in each protein we find that 16% of the proteins failed to find any overlap between the predicted spheres and the known catalytic cluster. It is important to understand why these failures occurred in order to improve the algorithm and assess whether there are specific types of enzyme on which the algorithm performs consistently badly.

Poor alignments

The alignments automatically generated by PSI-BLAST are the most likely point of failure. The optimal $E$-value cut-off for each family varies depending on its size and diversity. The single $E$-value cut-off used represents the best compromise, but still generates poor alignments for some families. To test whether poor alignments are the major source of error the difference between the conservation of the catalytic residues and the conservation of all residues was calculated and averaged for each group of results (correct, partial and incorrect), the results are shown in Figure 13. The different groups clearly show a variation in the distinction between conservation of catalytic and non-catalytic residues. In the correctly predicted group the difference is more than 0.3, this falls to 0.25 for the partially correct group, and the incorrect group has an average difference of only 0.15. Clearly, given the importance of conservation scores in making predictions, a lack of differentiation between the conservation of catalytic and non-catalytic residues will reduce the overall accuracy.

This trend implies that unusual conservation scores are responsible for a large part of the failure rate. The low difference in conservation scores in the failure group could be explained if these proteins all had low DOPS. The DOPS for each protein chain were averaged for each category and are also shown in Figure 13. There is a correlation between DOPS and the success of a prediction, however, looking at the scores themselves shows that, although some chains have very low DOPS, most are just as high as the average correctly predicted protein. If low DOPS were responsible for all of the failures then one would not expect the average conservation of the catalytic residues to vary across the three groups. However, a clear trend of increasing catalytic conservation in the correct predictions is detected and shown in Figure 13. How then to explain these anomalous conservation scores? The assumption must be that these enzymes are part of a larger family of proteins, which have different catalytic activities. Catalytic residues conserved within a sub-family would therefore vary between members of the family and not be necessarily conserved. Several examples of this can be seen in the failed structures. Calpain (1DKV) for instance contains an EF-hand domain, which is even found in non-enzymes. This means the catalytic residues of Calpain are not conserved in many of the homologues a PSI-BLAST search returns, whilst other residues involved in forming the EF-hand are conserved. This pattern of conservation is the inverse of what the network is expecting, and so it fails to correctly predict the catalytic residues.

Clustering errors

Of the 26 structures that failed to find the active site when only the top site was considered, ten also failed when all sites were considered. In these ten cases the error occurs prior to clustering, generally with poor alignments from PSI-BLAST. Of the remaining 16, 11 generated a lower scoring correct cluster and five generated a lower scoring partially correct cluster. These 16 cases are failures of the clustering algorithm to find the right cluster,
presumably because the signal from the true active site was weak compared to other sites in the protein.

If each structure is analysed by hand, the fault is generally obvious. The single-linkage algorithm is prone to forming long aspherical clusters, since two separate clusters can be joined even if only a single residue joins them. In several failures the true active site is a relatively compact cluster with a few high-scoring residues, whilst the top scoring prediction is a large cluster which out-scores the others by its size even if no single residue scores highly. Another problem is that the algorithm tends to select clusters buried in the protein, since these contain more residues than surface clusters, a human can easily spot that these are not suitable active sites.

Further work

Analysing the structures for high-scoring surface patches, as well as simple clusters might help in identifying the location of active sites, particularly if the top scoring cluster is deeply buried and hence unsuitable as a catalytic centre. Patch analysis has been used to identify other surface features, such as protein–protein interaction sites and ligand binding pockets.

The predicted clusters can also be used to automatically generate three-dimensional templates for analysis by one of the pattern searching algorithms, such as TESS and SPASM. Designing templates by hand is a time-consuming job and automated methods, such as this and the method recently described by Oldfield, could be used to quickly generate starting templates suitable for manual refinement.

The basic methodology of neural network scoring of residues and spatial clustering could be used to find other types of functional sites such as non-obligate protein–protein interfaces or protein–DNA interaction sites. Many of the secondary clusters found by this network, may be more confidently predicted by networks trained on these other functional classes. A novel protein structure could be presented to each network in turn and different types of functional sites identified at each stage.

Materials and Methods

Protein test set

The protein test set and the compilation of the data are described in detail in a recent paper by Bartlett et al. The original test set contains some proteins with homologous non-catalytic domains, for this study these redundant structures have been removed. The final test set contains 159 proteins from the PDB, containing no homologous pairs and covering all six top level enzyme classification (EC) numbers. This data set contains approximately 55,000 non-catalytic residues and 550 catalytic residues available for training the network.

Compilation of data

The catalytic residues were defined using the following rules:

1. Direct involvement in the catalytic mechanism (e.g. as a nucleophile).
2. Exerting an effect on another residue or water molecule, which is directly involved in the catalytic mechanism, which aids catalysis (e.g. by electrostatic or acid-base action).
4. Exerting an effect on a substrate or cofactor which aids catalysis, e.g. by polarising a bond which is to be broken. Includes ionic and electrostatic effects.

Note that residues that bind substrate, cofactor or metal ions are not included, unless they also perform one of the functions listed above.

Many studies have used the SITE records defined in PDB files as the basis for defining functional residues and sites. Unfortunately SITE records are not a homogenous data set, and there are no fixed rules on what may or may not be included in a SITE entry. Only 13 of the 159 PDB files in our data set contain SITE records, less than 10%. These 13 structures contain 50 catalytic residues, as defined above and 94 SITE residues. The overlap between these two groups contains 36 residues. We find therefore that in our data set 28% of catalytic residues are not found in the SITE records and only 38% of SITE residues are catalytic.

The following parameters were derived for each residue (catalytic and non-catalytic) in all 159 proteins:

- Conservation. The sequence of each chain in the protein was used to initiate a PSI-BLAST search of the NCBI non-redundant database (NRDB) with an E-value cut-off of 10−20 for inclusion in the next iteration. Each PSI-BLAST search was run to convergence or a maximum of 20 iterations. The final multiple alignment generated by PSI-BLAST was then scored for conservation and DOPS as described by Valdar et al.
- Relative Solvent Accessibility (RSA). NACCESS was used with standard parameters to calculate the RSA of each residue.
- Secondary structure. DSSP was used to extract the secondary structure for each residue. The DSSP classification was simplified to three categories: helix, sheet or coil/other.
- Cleft. Surfnet was used to define in which, if any, cleft the residue lay. If a residue lay in two or more clefts only the largest was recorded.
- Depth. The depth of a residue within the protein structure is defined as the average minimum distance between each of its atoms and the closest solvent accessible atom in the structure. NACCESS was used to define solvent accessibility.

Encoding and generation of data sets

Conservation, as calculated above, is already encoded as a suitably scaled factor between 0 and 1 (0 for no conservation and 1 for perfect conservation) and so is passed...
to the network as is. The RSA is a percentage and is scaled to between 0 and 1 before presentation to the network. Depth is scaled so that the deepest residue in each structure is scored 1 and surface residues 0.

The other parameters: residue type, secondary structure and cleft are categorical in nature, and are encoded using 1-of-C encoding. Amino acid type is encoded as an array of 20 inputs where one input is set to 1 and the rest to 0. Secondary structure is encoded by three input parameters. Cleft size is divided into four categories: no cleft, largest cleft, second or third largest cleft and fourth to ninth largest cleft.

An example encoding is shown in Figure 14 for a serine residue with conservation 0.7, DOPS score 0.9, depth 0.3, RSA 15%, in a coil region and lying in the largest cleft.

**Training the neural network**

The neural network software used is FFNN, a feed forward neural network trained using a scaled conjugate gradients algorithm. A single-layer architecture is used in all cases. In order to accurately measure the performance of the network it is trained using a ten fold cross validation experiment. The dataset is divided into ten equal subgroups, and then in each training run nine of the groups are used for training, whilst the network is tested on the single remaining group. The network is run ten times using a different subgroup as the test group each time. Here the dataset was divided by structure rather than residue, so each subgroup contains the data for approximately 16 structures. The ratio of catalytic to non-catalytic residues is approximately 1:60 in the training set. Presenting the data in this ratio causes the net to predict every residue as non-catalytic. The best balanced training set was found to have a ratio of 1:6. Each training group is balanced by discarding a random selection of the non-catalytic residues prior to training. Training was for 100 epochs, in every case the network converged to a stable error-level before training was terminated. The number of training epochs was not optimised, and in particular the performance of the test set was not used to optimise the stopping point in any way.

**Measuring performance**

In order to judge the neural network learning process, a suitable measure of performance is required. Total error (percentage of incorrect predictions) is not sufficient due to the highly unbalanced nature of the dataset. All of the statistics are derived from the following quantities:

- $p = \text{Number of correctly classified catalytic residues}$.
- $n = \text{Number of correctly classified non-catalytic residues}$.
- $o = \text{Number of non-catalytic residues incorrectly predicted to be catalytic (over-predictions)}$.
- $u = \text{Number of catalytic residues incorrectly predicted to be non-catalytic (under-predictions)}$.
- $t = \text{Total residues } (p + n + o + u)$.

The total error ($Q_{\text{total}}$) is given by equation (2):

$$Q_{\text{total}} = \frac{p + n}{t} \times 100$$  \hspace{1cm} (2)

To complement this, two other measures of performance are used, $Q_{\text{predicted}}$ measures the percentage of catalytic predictions that are correct and $Q_{\text{observed}}$ measures the percentage of catalytic residues that are correctly predicted. The formulae for these two parameters are shown in equations (3) and (4):

$$Q_{\text{predicted}} = \frac{p}{p + o} \times 100$$  \hspace{1cm} (3)

$$Q_{\text{observed}} = \frac{p}{p + u} \times 100$$  \hspace{1cm} (4)

A measure of performance that takes both these factors into account is the MCC. The formula for calculating MCC is shown in equation (5):

$$\text{MCC} = \frac{pm - ou}{\sqrt{(p + o)(p + u)(n + o)(n + u)}}$$  \hspace{1cm} (5)

**Ranking and clustering**

The residues in each structure are ranked by network score, and all residues scoring above a cut-off value are used in the clustering algorithm. A pair of residues are clustered together if any of their atoms lies within 4 Å of each other. Each cluster is then defined as a sphere with its centre at the geometric centroid of all the Cα atoms of the component residues (Cα for glycine) and a radius such that all the Cα atoms lie within the sphere. The first ranking cut-off was set at 35% of the highest scoring residue. If any sphere in a structure had a radius greater than 15 Å, the clustering was repeated, increasing the ranking cut-off by 1% until no sphere was greater than 15 Å in radius. Single residue clusters were discarded at this stage.

The definition of the known sites is the same. Spheres were defined for each active site with centres at the centroid of the Cα atoms and radii such that all the Cα atoms are within the sphere. Proteins with single catalytic residues were set a radius of 3 Å.

**Acknowledgements**

Thanks to Dr Adrian Shepherd at UCL and Dr Craig Porter at the EBI for useful discussions on neural networks, clustering algorithms and other topics. Thanks to the Medical Research Council for financial support. G.J.B. was supported by a BBSRC CASE studentship in association with Roche Products Ltd.

**References**


Edited by B. Honig

(Received 27 November 2002; received in revised form 26 February 2003; accepted 2 April 2003)