Using microenvironments to identify allosteric binding sites.

Christopher E. Foley[1,2], Sana Al Azwari[1], Mark Dufton[2] and John N.Wilson[1]
[1]*Dept of Computer & Information Sciences,* [2]*Dept of Pure & Applied Chemistry, University of Strathclyde, Glasgow, UK.*

***Abstract*—Protein amino acid residues can be classified by their chemical properties and data mining can be used to make predictions about their structure and function. However, the properties of the surrounding residues contribute to the overall chemical context. This paper defines microenvironments as the spherical volume around a point in space and uses these volumes to determine average properties of the encompassed residues. The approach to index generation rapidly constructs microenvironment data. The averaged chemical properties are then employed in allosteric site prediction using support vector machines and neural networks. The results show that index generation performs best when microenvironment radius matches the granularity of the index and that microenvironments improve the classification accuracy.**

I. Introduction

High throughput screening makes a useful contribution to drug discovery. During this automated process, every molecule in a large compound bank is individually tested for activity against a drug target, usually a protein. The leads (the best matching compounds) are taken to the next stage where their structure is refined to meet other biochemical requirements. If the combined protein and lead can be crystallised, it may be possible to elucidate the three dimensional molecular structure and thereby determine the location on the protein at which the lead molecule is bound. If this site is remote from an active site in the protein, then the molecule is said to be an allosteric regulator. Drugs based on allosteric regulators are important because of their specificity and potential for lower toxicity. Unfortunately, most of the active compounds turned up by high throughput screening are not allosteric regulators.

Proteins are made by the polymerisation of amino acids into a linear chain that is folded into a three dimensional (3D) structure. The folding pattern brings the functional parts of the protein together and adjusts its configuration in response to binding interactions. Much of the functionality of a protein depends on small positional adjustments. We define a microenvironment as the localised three dimensional spherical neighbourhood surrounding a particular point within the space of an object. In this case the points are represented by the α-carbons within a protein structure. The microenvironment encloses a variable number of atoms from the protein, depending on the radius of the sphere.

The mean temperature factor of the enclosed atoms is a measure of local flexibility around the microenvironments central α-carbon. This is a consequence of the dependence of protein fold adjustment on the plasticity of a local topology. Since any individual atom may be part of several spheres this approach provides an estimate of the protein's behaviour in the surrounding area rather than behaviour at the point represented by each atom. This is useful because the activity of a protein is influenced by the general topology rather than point-by-point parameter values.

Support vector machines (SVM) and neural networks (NN) span the range of prediction accuracies for establishing classifications in datasets [1]. An SVM [2] is a supervised learning mechanism that generates a hyperplane separating data in a training set. SVMs have been used in bioinformatic research to generate optimal classifications of sites on protein chains [3]. Artificial NNs represent an alternative approach to classifying input data. SVMs and NNs have been used to distinguish the likely positions of protein-protein interaction sites in protein chains [4] and offer the potential for providing the same service in distinguishing allosteric sites.

It is not feasible to pre-compute all the combinations for extensive collections of protein data. Generating microenvironments on-the-fly provides sufficient flexibility and at the same time can support rapid exploration of data. Three dimensional (3D) grid methods have long been known to provide a basis for accelerating the performance of processing spatial data [5]. However in the scenario of varying the level of abstraction of microenvironment data, the most appropriate dimensions for grid indexes are uncertain.

The contribution of the work described in this paper is to identify the best way of generating an on-the-fly index for the rapid association of atoms within a protein with the centres that contain these atoms. This methodology is then used to demonstrate that the assembly of clustered data makes a significant contribution to predicting the localisation of allosteric sites. The rest of the paper is organised as follows: Section II establishes the context of related approaches. Section III describes index generation and its use in the microenvironment assembly algorithm and Section III presents the experimental work, the results of which are presented in Section IV. The paper concludes with an evaluation of the results and the potential for further work.

II. Related work

Early recognition of the power of quantising the space of individual molecules came from Leventhal [5]. This approach was further contextualised by Bentley [6] who assumed a quantisation based on search radius. The approach

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

described in the current work explores the assumption that the optimum cell size of the quantisation is the same as search radius. Establishing the optimal approach is an essential step in providing a suitably efficient method of microenvironment assembly.

Fixed size microenvironments have been used as a basis for k-means clustering with a view to exploring protein structure [7]. This approach has also been successful in identifying calcium binding sites [8]. Predicting allosteric hotspots on the basis of physicochemical parameters of amino acids overcomes some of the problems of high throughput screening. SVMs have been successfully used in this context to classify allosterically active sites [9], [10]. Limited availablitiy of data that identifies allosteric sites has hitherto been a problem. The latter of these studies based SVM training on 44 hotspots (residues with empirically or prediced allosteric involvement) and 50 non hotspots (residues that had been demonstrated not to contribute to allosteric activity). Data available in the Allosteric Site Database [11] (ASD) provides a basis of a larger-scale study of this approach.

Prediction of allosteric sites has focused on the use of point data associated with residues in protein chains. The work described in this paper confirms the efficiency of grid indices for the generation of spherical localisations centred on α-carbon atoms in a protein chain. The spheres provide a more representative value for each region and result in improved prediction accuracy when applied to the classification of allosteric activity in the residues that make up a protein chain. Furthermore, the prediction is based on the distinction between hotspots and unknowns, thereby circumventing the need to identify non-hotspots and consequently providing potential for a scalable solution.

III. Methods

Microenvironment assembly determines the atoms that lie inside the sphere that is centred on each α-carbon in a protein chain. The simple approach of calculating the Euclidean distance between all amino acid/atom pairs is inefficient because of the large number of potential pairs. Cell partitioning [6] is used to pre-organise data so that only nearby atoms are considered as candidates during microenvironment assembly. Generation of the cell index is carried out on-the-fly having previously established the overall maxima and minima of spatial coordinates in the PDB collection and used this to produce a 3D grid. The coordinates of each atom associate it with a particular cell in this grid. The index identifies a candidate set of atoms that can be formed from the surrounding cells, immediately ruling out distant atoms from consideration. Only atoms within a reasonable distance of the sphere centre become candidates. The distances between the candidates and the sphere centre are calculated and the appropriate atoms included in the sphere. The index can be tuned by altering the size of the 3D cells and Figure 1 shows how the candidate set is narrowed down by choosing only the cells that intersect with the sphere. Given the importance of this step in on-the-fly assembly of microenvironments, it is necessary to assess whether the optimal cell edge length is the same as radius size or whether a sub-multiple (L/n) of radius size would be more appropriate.

At one extreme, a single large cell will place all the atoms together, effectively removing any benefit from the index. At the other end of the spectrum, if the cell size is too small each atom will have its own box, negating the advantage. Figure 1 suggests that choosing a cell size that matches the sphere radius constrains the candidate space but that sub-multiples ($L/2$, $L/3$ etc.) might be more effective.

Experimental work was carried out to evaluate the optimal approach to indexing the Protein Data Bank collection [12] with a view to rapid assembly of microenviroments. A second set of experiments evaluates the effect of physicochemical parameter variation on classification of micronenvironment centres. Experiments were run to configure the 3D grid index in the context of the collected data structures from the PDB. Microenvironments were then assembled using varied sphere radii and protein sizes in order to allow a comparison. Experiments were conducted on a 3 GHz Intel Pentium 4 processor with 1 GB RAM, running Zenwalk Linux 6.2. The algorithms were implemented in Java 6.

To obtain a representative test dataset, the protein chains present in the PDB were divided into groups by chain length (1–50 amino acids in the first group, 51–100 in the second, etc.) and one chain was chosen at random from each group. The final dataset is shown in Table I.

The performance evaluation of microenvironment assembly was carried out by repeating the algorithm 1000 times. In order to make sure the compiled and optimised execution was measured, the 1000 measurements were repeated until two consecutive measurements were within 10% of each other.

To determine the best cell size, the protein size was kept constant. Chain E from 1ZPU was chosen, which fixed the number of amino acids at 529.

An index using cells that are too small will take a long time to create while very large cells will approach $O(n^2)$ in terms of microenvironment assembly performance. Somewhere between these two extremes must lie the maximum efficiency. The experiment was run at sphere radii of 4, 5, 6,

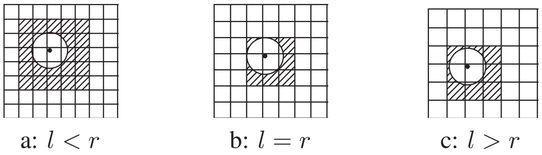
a: $l < r$ b: $l = r$ c: $l > r$

Figure 1: Variation of sphere radius (r) and cell length (l).

ID	Ch.	Len	ID	Ch.	Len	ID	Ch.	Len	ID	Ch.	Len
1D9M	A	18	1AZP	A	66	2I8T	B	149	2IX3	B	972
1BGL	F	1021	3DEE	A	197	1J2Q	B	223	2QPQ	C	296
2QN1	A	813	2VC9	A	882	1MIQ	B	327	2OF6	B	400
1JRP	G	450	1UYT	A	681	1N7O	A	721	2HLD	S	480
1ZPU	E	529	1EFK	A	553	2PPB	M	1119	1WZ2	B	948
2AHX	B	615	1JRP	B	760						

Table I: Protein chains used for algorithms benchmarking.

Parameter	Characteristic
Temperature factor (*B*-factor)	The flexibility of the protein at a particular atom.
Druggability	The likelihood that a residue will be targeted by a drug-like molecule.
Hydrophobicity	The potential for folding at a particular residue
Total atomic weight	The local size around a residue.
Residue number	The position of a residue in the protein chain.

Table II: Dataset parameters

ID	Ch.	ID	Ch.	ID	Ch.	ID	Ch.	ID	Ch.
1ABB	A	1FTA	A	11Q	A	1KFL	A	1NTK	C
1NXG	A	OG2	A	1PFK	A	1Q0B	A	1S9I	A
1T48	A	1T5A	A	2BRL	A	2BU2	A	BXD	A
2D5Z	A	2I80	A	2JC9	A	R1R	A	2VK1	A
3BEO	A	3C1N	A	I0R	A	3KCC	A		

Table III: Sample dataset set for classifying allosteric sites.

	Total residues		Allosteric residues	
	Train	Test	Train	Test
ASD native	5196	5156	178	188
ASD resamp	2214	5156	536	188
Enhnce ASD resamp	2816	5156	1230	421
Enhnce random resamp	3009	5156	1492	565

Table IV: Dataset cardinality

7, 8, 9 and 10Å. For each sphere size, the cell size was varied from 4 Å to 20 Å in steps of 1Å. The best cell size was deduced from the above experiments and used to benchmark the cell index at sphere radii of 4, 5, 6, 7, 8, 9 and 10Å for each chain length in the dataset.

Having validated the approach to generating microenvironments it was possible to use the method to explore the efficacy of sphere-based predictions of allosteric involvement.

A set of proteins with an indication of their allosteric residues was obtained from the Allosteric Site Database [11]. For each protein in this collection, a single PDB identifier was selected. To counteract the sparsity of recognised allosteric sites in the collected protein structures, the experimental set was restricted to the 24 chains with the largest set of allosteric residues (Table III).

The mean for each of temperature factor, hydrophobicity, druggability, total atomic weight and residue number were assembled for each microenvironment within radii of 0Å-50Å (stepped at 10Å) of the central amino acid. These parameters have been shown to be useful in distinguishing residues that contribute to protein-protein interfaces [13] and in addition address orthogonal characteristics (Table II). The assembled microenvironments, together with identification of the residues that contributed to allosteric sites produced a set of vectors suitable for input into classifiers. LIBSVM [3] and Matlab were used to explore the potential for SVM and NN methods in classifying residue contribution to allostery within the specified protein data set.

Table IV shows that there were a total of 366 allosteric hotspot residues in the sample data collection, which consisted of more than 10000 residues. Such sparse datasets provide a challenge for classification methodologies. To counteract this problem, residue sites that are close neighbours of those allosteric residues identified in the ASD were also included as classification targets. This enhancement reflects the influence that such near neighbours have on allosteric site. In addition, in some cases, the ASD source data identifies uncertainty over which residue of a pair of near neighbours forms the allosteric site. Near neighbour inclusion was implemented by adding those residues within 4Å of an allosteric site into the set of target residues.

Alternative approaches that can mitigate sparse datasets include combined over- and under-sampling [14]. Following this approach, the assembled microenvironment data was randomly separated into non-overlapping training and test sets. Random replication of the minority (positive) cases and sampling of the majority cases was applied only to the training set. The test set was used without further replication or sampling. These enrichment processes produced four data sets: *ASD native*: the data recovered from the Allosteric Site Database; *ASD resample*: the ASD expanded by resampling the positive residues and randomly sampling the negative residues both by a factor of 3; *Enhanced ASD resample*: the ASD sites with near-neighbours added into the positive residue set and resampled as in ASD resample; *Enhanced random resample*: a further set in which a random sample of similar size to the set of residues identified in ASD were marked as positive residues. This was enhanced and resampled as described above (Table IV).

IV. Experimental results and discussion

The maximum number of amino acids in a single chain in the PDB is 4128, suggesting that index generation will be around 0.6 s in the worst case. It can be seen from Figure 2, that the most efficient cell size is equal to the sphere radius. As the cell size increases from this global minimum, the time for the algorithm to run increases steadily. This is consistent with the larger cells holding progressively more atoms and therefore requiring more distance calculations. As the cell size decreases from the global minimum, the trend is for the time to increase. This is because more cells are required and their creation becomes the most time-intensive step. However Figure 2 also shows local minima at half the optimum cell size. Consider determining the sphere at a 7Å radius. When the cell size is also 7Å, the candidate list is drawn from the central cell and all of the surrounding cells. If the cell size is decreased to 6Å the central cell and the surrounding ones still have to be checked. However, now

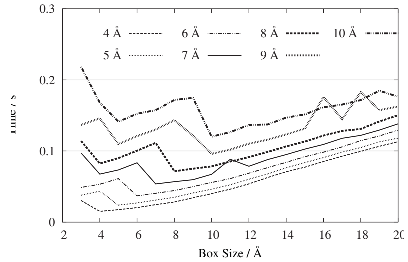

Figure 2: Effect of cell size on execution time for different sphere radii.

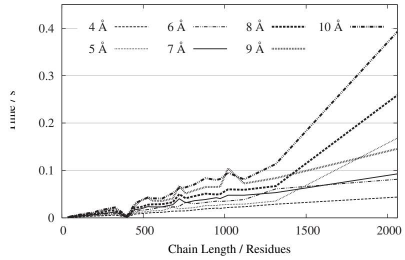

Figure 3: Index generation at different sphere radii (cell size).

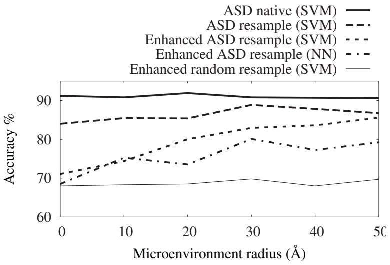

Figure 4: Accuracy at varying microenvironment radii.

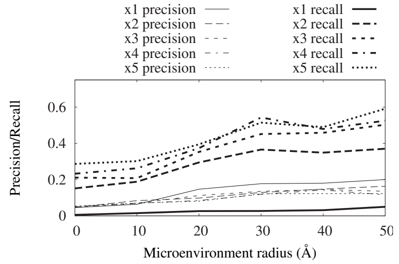

Figure 5: Precision and recall at varying microenvironment radii.

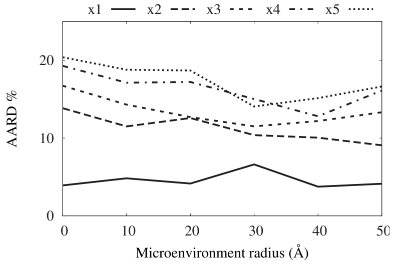

Figure 6: AARD% at varying microenvironment radii.

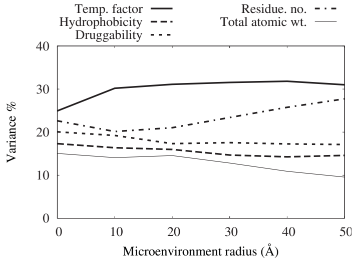

Figure 7: Parameter impact. (ASD native)

the range of the sphere can include atoms up to two cells away. If we go below 3.5Å (half the optimum box size), we have to consider atoms three cells away. One could expect another local minimum at 1.75Å, another at half this and so on (Figure 3). The experimental work verifies the assumption that the most appropriate cell length matches sphere radius.

The effect of varying the radius of the microenvironment on the accuracy of SVM and NN predictions of involvement in allosteric activity is shown in Figure 4. In the case of the ASD native data set, the apparent accuracy is a consequence of the excess of negative cases. The test set is overwhelmingly categorised as negative instances but both precision and recall are limited. The accuracy of the enhanced set is lower than the resampled set but in both cases there is a marked improvement up to a sphere radius of about 20Å with further small improvements therafter. This effect is noticable in results both from SVM and NN classifiers. The precision and recall of the enhanced ASD data set (Figure 5) vary with the level of resampling although beyond x3 (ie resampling the hotspots three times and sampling a third of the other residues) there is limited improvement. The effect of increasing microenvironment radius is detectible at all levels of resampling. The limited recall of the resampled dataset is a consequence of overfitting the separating hyperplane in the context of an excess of negative cases in the test dataset (4968 negatives and 188 positives shown in Table IV). This is to some extent mitigated in the enhanced dataset. The enhanced random resampled data set shows that varying the microenvironment radius has no impact in classifying randomly chosen residues. This indicates that the SVM is successfully classifying residues that contribute to the allosteric site in the enhanced data set.

The experimental work verifies the assumption that the most appropriate cell length matches sphere radius. This result provides confidence in the approach to optimal performance of microenvironment assembly, which is necessary for locating good classifiers within the search space. The on-the-fly approach is efficient enough to remove the necessity for materialising microenvironments. This improves the utility of the method since it removes the need for predicting the combinations of parameters that a drug researcher will find useful. If a user interface must respond to a mouse click or a keystroke within 0.1 s, the 3D grid index continues to meet the criteria up to about 1200 amino acids for the higher sphere sizes and over 2000 amino acids for sphere sizes of 7Å and under. This makes it feasible to build a direct manipulation interface to large data collections such as the PDB and provides support for interactive data mining. The experiments show that the performance of the index is dependent on the sphere size, with larger radii making the index less efficient. The radii chosen for the experiment were relatively small compared with the total space occupied by the data points. In the course of the experiments, it was assumed that the data points were distributed roughly evenly. The 3D grid index has not been tested with other distributions though it is likely that it will still provide a performance increase.

The effect of increasing radius of microenvironment is generally to improve the accuracy of predications of allosteric activity as indicated in Figure 12 particularly by the SVM and NN classifications for the enhanced ASD resample data. The paucity of confirmed allosteric sites among the residues of the protein data set used, necessitates the enhancement of this data and its resampling. Demerdash et al [10] show precision and recall as 55%-67% and 68%-92% using SVMs to classify a smaller data set. These precision and recall values are based on training and test sets consisting of residues that are classified as either hotspots or non-hotspots depending on experimental evidence of the activity of each residue included. Clearly, this approach to identifying non-hotspots is costly and not appropriate for applications that aim to classify large collections of data such as that available in the PDB. By contrast, the approach presented in this paper seeks to classify data that is categorised as either a hotspot or an unknown. The returned precision and recall are lower (20% and 60% respectively) but the technique will scale more effectively since it is based on unknowns rather than non-hotspots.

The absolute average relative deviation (AARD) (Figure 6) also shows the beneficial effect of increasing microenvironment radius, particularly in the context of the enhanced dataset. Re-interpreting the data presented by Demerdash et al gives AARD% in the range 25%-34% for the most successful models. The larger data set and use of unknown residues rather than non-hotspots in the current study returns an AARD of 10-15% for resampled training sets at the most effective microenvironment radius. Other studies on prediction of allosteric sites [9] have produced AARD values of 6%-7% but this is in the context of training and test sets based on variants of a single molecule rather than the range of different proteins used in the current experiments. The AARD% in the current study suggest that the scaling that becomes possible when training sets consist of hotspots and unknowns rather than hotspots and non-hotspots is not acheived at the expense of the utility of the prediction. Figures 7 characterise the variability of the data in the training and test sets. The ratio of the total variance provided by each component changes as a consequence of increasing sphere radius with temperature factor consistently providing the largest component. Each parameter appears to provide a sizable component of the total variability, indicating that each is measuring an orthogonal characteristic.

V. Conclusion

The experimental work reported has evaluated the efficiency of a parameterised 3D grid index for generating microenvironment data for use in the classification of amino acids in terms of their contribution to protein-protein interface sites. The index was evaluated with protein atomic coordinates and has been shown to be most efficient when the cell size matches the granularity of the summary. An underlying assumption is that the dataset is approximately evenly distributed and that the sphere size is small relative to the space occupied by the data set as a whole. Despite these assumptions, the algorithm may be effective for larger spheres and different distributions.

The impact of this approach to characterising localities in a protein is shown to improve the utility of predictions of allosteric activity. In the context of the dearth of hotspots (residues that are empirically determined to contribute to allosteric activity), enhancing and resampling the training data seems to provide a way of generating suitable SVM models. In comparison with previous work in this area, the approach presents a scalable method of predicting allosteric activity in large datasets.

References

[1] J. Bisbal, G. Engelbrecht, M. Villa-Uriol, and A. Frangi, “Prediction of cerebral aneurysm rupture,” in *Proc. DEXA*, 2011, vol. 6861, pp. 59–73.

[2] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[3] C. Chang and C. Lin, “LIBSVM: A library for support vector machines,” *ACM TOIST*, vol. 2, pp. 27:1–27:27, 2011.

[4] R. Liu, W. Jiang, and Y. Zhou, “Identifying protein-protein interaction sites in transient complexes,” *Amino Acids*, vol. 38, no. 1, pp. 263–270, 2010.

[5] C. Levinthal, “Molecular model-building by computer,” *Scientific American*, vol. 214, pp. 42–52, 1966.

[6] J. Bentley, D. Stanat, and E. Hollins Williams, “The complexity of finding fixed-radius near neighbors,” *Information Processing Letters*, vol. 6, no. 6, pp. 209–212, 1977.

[7] S. Wu, T. Liu, and R. Altman, “Identification of recurring protein structure microenvironments,” *BMC Struct Biol*, vol. 10, no. 4, 2010.

[8] S. Bagley and R. Altman, “Characterizing the microenvironment surrounding protein sites,” *Protein Science*, vol. 4, pp. 622–635, 1995.

[9] E. Pourbasheer, S. Riahi, M. Ganjali, and P. Norouzi, “QSAR study of c allosteric binding site of HCV NS5B polymerase inhibitors,” *Molecular Diversity*, vol. 15, pp. 645–653, 2011.

[10] O. N. A. Demerdash, M. D. Daily, and J. C. Mitchell, “Structure-based predictive models for allosteric hot spots,” *PLoS Comput Biol*, vol. 5, no. 10, p. e1000531, 10 2009.

[11] Z. Huang *et al.*, “ASD: a comprehensive database of allosteric proteins and modulators,” *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D663–D669, 2011.

[12] H. M. Berman *et al.*, “The Protein Data Bank,” *Acta Crystallogr. D*, vol. 58, no. 6 Part 1, pp. 899–907, 2002.

[13] I. Ezkurdia *et al.*, “Progress and challenges in predicting protein-protein interaction sites,” *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 233–246, 2009.

[14] C. X. Ling and C. Li, “Data mining for direct marketing: Problems and solutions,” in *KDD*, 1998, pp. 73–79.