國立交通大學

生物資訊研究所

碩士論文

由蛋白質三級結構預測其蛋白質交互作用區域

Prediction of Protein-Protein Interaction Sites from Three-Dimensional Structure

研 究 生：黃存操

指導教授：黃鎮剛　教授

中華民國九十四年六月

由蛋白質三級結構預測其蛋白質交互作用區域

Prediction of Protein-Protein Interaction Sites from Three-Dimensional Structure

研究生：黃存操	Student：Tsun-Tsao Huang
指導教授：黃鎮剛	Advisor：Jenn-Kang Hwang

國立交通大學
生物資訊所
碩士論文

A Thesis
Submitted to Institute of Bioinformatics
College of Biological Science and Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Bioinformatics

June 2005
Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

由蛋白質三級結構預測其蛋白質交互作用區域

學生：黃存操　　　　　　　　　　指導教授：黃鎮剛教授

國立交通大學生物資訊研究所碩士班

摘　　要

理解蛋白質交互作用的機制，對於分析蛋白質功能或是了解蛋白質交互作用網路有著舉足輕重的角色。因此，若能預測蛋白質交互作用的區域(protein-protein interaction sites)，對於解讀蛋白質功能與其反應將很有幫助。在這篇論文中，我們使用機器學習(machine-learning approach)的方法，利用胺基酸組成、蛋白質二級結構、胺基酸相對暴露率(relative solvent accessible surface area)、演化資訊與蛋白質結構等資訊，從蛋白質的三級結構來預測此蛋白質與其他蛋白質發生交互作用的區域。我們在一個蛋白質基準集合上，測試了我們的方法，並且順利地預測了蛋白質交互作用區域。這結果顯示出我們的方法在預測蛋白質交互作用區域上是有用的。

Prediction of Protein-Protein Interaction Sites from Three-Dimensional Structure

Student：Tsun-Tsao Huang　　Advisor：Dr. Jenn-Kang Hwang

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

The knowledge of protein-protein interactions is essential for the understanding of protein functions and protein-protein interaction networks. Hence, the capability to identify protein-protein interaction sites is crucial to decipher the reaction mechanisms of protein function.

In this work we develop am approach based on machine-learning method to predict protein-protein interaction sites from three-dimensional structure. We have tried a multiple of feature vectors such as amino acid composition, secondary structure, relative solvent accessible surface area, position substitution specific matrix and structural neighboring residues. Our results compare favorably with those of others for a benchmark dataset.

誌謝

這篇論文要感謝我的指導教授黃鎮剛老師，提供良好的研究環境與設備對研究助益良多。還有父母親在經濟上面的幫助，讓我可以無後顧之憂的做研究。此外，感謝尤禎祥學長辛勞地管理機器們，林勇欣學長在實驗上面的建議，游景盛學長指導如何操作支持向量機器，還有羅淵仁學長帶大家出去逛逛，放鬆心情。與梁涵堃學長之間的討論，因緣巧合地讓我做了這個題目。與施建華同學、林肇基同學、陳啟德同學之間的討論與交流，讓這篇論文得以順利的完成。

目　　錄

inhibitor colored lightblue and is represented in mesh and its partner is colored red and is represented in line. Blue region is interface and green region is predicted as interface. 1avwB

Figure 16. Soybean trypsin inhibitor from Soybean (*Glycine max*). Soybean trypsin inhibitor is colored lightblue and is represented in mesh. Violet region denotes underprediction (false negative) and blue region is predicted correctly (true positive). Yellow region denotes overprediction (false positive) and green region is predicted correctly (true positive).

Figure 17. Pancreatic trypsin inhibitor, BPTI from Cow (*Bos taurus*). BPTI is colored lightblue and is represented in mesh and its partner is colored red and is represented in line. Blue region is interface and green region is predicted as interface. 1bthP

Figure 18. Pancreatic trypsin inhibitor, BPTI from Cow (*Bos taurus*). BPTI is colored lightblue and is represented in mesh. Violet region denotes underprediction (false negative) and blue region is predicted correctly (true positive). Yellow region denotes overprediction (false positive) and green region is predicted correctly (true positive).

Figure 19. Ovomucoid domains from Turkey (*Meleagris gallopavo*). This protein is colored lightblue and is represented in mesh and its partner is colored red and is represented in line. Blue region is interface and green region is predicted as interface.

Figure 20. Ovomucoid domains from Turkey (*Meleagris gallopavo*). This protein is colored lightblue and is represented in mesh. Violet region denotes underprediction (false negative) and blue region is predicted correctly (true positive). Yellow region denotes overprediction (false positive) and green region is predicted correctly (true positive).

Figure 21. beta2-microglobulin from Human (*Homo sapiens*). This protein is colored lightblue and is represented in mesh and its partner is colored red and is represented in line. Blue region is interface and green region is predicted as interface. 1ao7B

Figure 22. beta2-microglobulin from Human (*Homo sapiens*). This protein is colored lightblue and is represented in mesh. Violet region denotes underprediction (false negative) and blue region is predicted correctly (true positive). Yellow region denotes overprediction (false positive) and green region is predicted correctly (true positive).

Figure 23. Hirudin from Leech (*Hirudo medicinalis*). Hirudin is colored lightblue and is represented in mesh and its partner is colored red and is represented in line. Blue region is interface and green region is predicted as interface. 4htcI

Figure 24. Hirudin from Leech (*Hirudo medicinalis*). Hirudin is colored lightblue and is represented in mesh. Violet region denotes underprediction (false negative) and blue region is predicted correctly (true positive). Yellow region denotes overprediction (false positive) and green region is predicted correctly (true positive).

INTRODUCTION

Protein-protein interactions are important in numerous biological processes such as immune response, enzyme catalysis, and signal transduction.[1] The discovery of interaction sites is useful for the prediction of unknown protein-protein interactions, for the drug design as targets in proteomics[2], for the library design of phage display and for limiting search space for docking studies[3]. Although prediction of interaction sites with protein complexation is important, it remains a challenge in molecular biophysics.[4]

Protein-protein interaction sites can be deduced by experimental and computational methods. Experimental methods can be grouped into several categories. Analyzing protein complexes structures[5] is the most direct method. Analyzing structures requires atom coordinates of protein complexes structures, but it is generally more difficult to obtain the structure of protein complexes than that of protein monomer structures.[6] Phage display[7-9] and mutagenesis[10-12] have also been attempted to discover interaction sites. These experimental methods usually lead to relatively high accuracy, but most of them are time-consuming and cost-expensive.[2]

By using patch analysis on six chemico-physical properties (salvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area) from protein structures, Jones and Thornton obtain 66% accuracy in protein level.[13] Gallet et al.,[14] detected DNA-binding sites (95% detected) and Ca-binding domains (85% detected) by exploiting the difference of hydrophobicity distribution in linear stretches of sequences of interface residues and of non-interface residues. But Fariselli et al.[15] confirmed that the chemico-physical properties of interaction residues are difficult to distinguish from protein surface residue. Fariselli et al.[15] used neural network with sequence profile from HSSP[16] and obtained 73%

prediction accuracy in three-fold cross-validation. Zhou et al.[1] included sequence profile and residue neighbor list and used neural network to obtain 70% prediction accuracy. In addition to sequence profile information, Liang et al.[6] analyzed 16 protein chains and showed that the residue with the highest energy score on the surface of a small protomer is very possibly the key interaction residue. Gao et al.[17] achieved 88% success rate on 75 hot-spot residues from their structure-based method.

It is difficult to judge the importance of features or the advantage of prediction methods in protein-protein interaction sites prediction because different features, different methods and different validation methods are used previously.[1,3,6,13,15,17-21] Different validation methods are used and find a looked better result because prediction of protein-protein interaction sites is difficult. The difficulty of protein-protein interaction sites prediction comes from the large diversity of protein-protein interactions.[18]

The support vector machine (SVM) method[22] has recently become popular in computational biology.[23-27] It is successful in applying the SVM based on multiple feature vectors on protein fold assignment[28] and subcellular localization prediction.[24] In addition, knowledge of a protein's tertiary structure is a prerequisite for the proper understanding of its function.[29] It is assumed that using information from structurally homologous structures on prediction is more potential than using that from non-homologous structures. In this work, the SVM is used to predict the protein-protein interaction sites based on multiple feature vectors from structural homologous structures and from non-homologous structures in order to prove the aforementioned hypothesis.

METHODS

Support Vector Machine

The SVM[22] tries to find the separating hyperplane with the largest distance between two classes, measured along a line perpendicular to this hyperplane. However, in practice, these data to be classified may not be linearly separable. To overcome this difficulty, SVM non-linearly transforms the original input space into a higher dimensional space, it is possible that data can be linearly separated. In the training process, only part of the training data are used to construct the hyperplane, hence avoiding the overfitting problem usually plaguing other machine learning methods. These data constructing the classifier are called support vectors. Preliminary tests show that the radial basis function (RBF) kernel gives results better than other kernels. Therefore, in this work we use the RBF kernel for all the experiments.

An important issue of optimizing SVM results is to deal with data unbalance. We avoided data unbalance by adjusting the weight of penalty parameter. The reciprocal of the data number of each class is used as its penalty parameters. In this work, all SVM calculations are performed by using LIBSVM[30], a general library for support vector classification and regression.

Coding schemes

Previous work[3] shows that protein descriptors based on the generalized *n*-peptide compositions are effective in protein-protein interaction sites prediction. If $n = 1$, then the *n*-peptide composition reduces to the amino acid composition, and if $n = 2$, the *n*-peptide composition gives dipeptide composition. When n gets larger, the *n*-peptide compositions will cover more global sequence information, but at the same time, such a coding scheme becomes not only impractical from a computational viewpoint but also undoable from a learning viewpoint. However, the size problem can be overcome if we regroup the amino acids into smaller groups of classes according to their physicochemical properties or structural properties. Previous work shows that 9-peptide composition with one target residue and four flanking residues on both sides of target residue is better in protein-protein interaction sites prediction[3] so 9-peptide composition is used in this work. Notation *AA* denotes the 9-peptide composition of amino acids. Figure 1 shows coding scheme *AA*. Notation *SS* denotes secondary structure information of 9-peptide composition of amino acids. Figure 2 shows coding scheme *SS*. Notation *RSA* denotes relative solvent accessibility[31] information of 9-peptide composition. Figure 3 shows coding scheme *RSA*. Profile element values in PSSM are rescaled by Kim's function[26] given by

$$f(x) = \begin{cases} 0 & \text{if } x \le -5 \\ 0.5 + 0.1x & \text{if } 5 > \text{x} > 5 \\ 1 & \text{if x} \ge 5 \end{cases} \quad . \tag{1}$$

Notation *PSSM* denotes above PSSM information of 9-peptide composition and is showed by figure 4. Notation *PSSM'* denotes PSSM information (rescaled by sigmoid function) of 9-peptide composition. Coding scheme *PSSM* is work in secondary structure prediction.[26] We use the notation *3D* to denotes information from 9 nearest residues in three-dimensional structure. The notation *3D-AA* and *3D-PSSM* (see figure 5) are used when these 10 residues (one target residue and nine neighboring residues) are encoded by their amino acid composition and PSSM information, respectively. Similar sequence and structure coding schemes has also been successfully applied to interaction sites prediction[19] and to metal-binding sites prediction[32].

SVM training and testing

Criteria of correctness in prediction

Each residue of structures is either exposed residue or buried residue. The solvent accessibility (*ACC*) of residues from DSSP[33] are used to determine residues are exposed or not. A residue is exposed if its relative solvent accessibility (*RACC* = *ACC/MaxACC*, with maximal accessibility for the amino acids)[31] is larger than the threshold *RACC*. Threshold *RACC* is 25% used by Yan et al.[3] and just these exposed residues (also called surface residues) are used in experiment. Each exposed residue is either in interface or not in interface. The state of an exposed residue (in interface or not) is demanded in training SVM and in validation of prediction results. An exposed residue is defined to be an interface residue if its *ACC* in the complex is less than its *ACC* in the monomer by at least 1 Å^2.[3] These procedures are showed in figure 6.

SVM Training schemes with and without structural homology information

Two different schemes are used in SVM to build training dataset. The first method is, for each testing protein chain in data set, the rest protein chains are used as its training dataset. Because the sequence identities between sequences is data set are below 30%, there is no homology between training dataset and testing protein chain. This training scheme is without structural homology information. D_n denotes this training scheme showed in figure 8.

Structurally homologous structures are collected as training datasets. Figure 7 shows the collection strategy. If query structure is in SCOP[34], entries except query with the same superfamily label in SCOP are collected as query's training dataset. Superfamily level are used in collection for the reason that proteins in the same superfamily are suggested a common evolutionary origin.[35] If query structure does not belong to any superfamily in SCOP, program ALIGN[36] is used to search the most similar structure in SCOP. Entries with the same superfamily label of this similar structure in SCOP are collected as training dataset. If the sequence identity calculated by ALIGN between query structure and the most similar structure is below 30%, program CE[37] is executed to search the most similar structure in SCOP. Entries belonging to the superfamily of this most similar structure are used as training dataset. If the query structure is composed of two chains, results of each individual chain are merged to a new result for this structure. D_h denotes above training scheme with structural homology information. Figure 9 is the flow chart of training scheme D_h. The testing dataset of my method is different from the training dataset of mine. When searching for structural homologous structures, CE and ALIGN cannot always find out reasonable similar structures in SCOP alone, so previous procedures are necessary to collect the structural homology training datasets. Figure 9 shows the flow of our prediction.

Evaluation of Predictive Performances

Predictive performance is assessed by sensitivity (*sens*), specificity (*spec*), accuracy (*accu*) and Matthews Correlation Coefficient (*MCC*)[38]. The overall prediction accuracy (*accu*) is given by:

$$accu = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (2)$$

where *TP*, *FP*, *TN*, *FN* are the numbers of true positives, false positives, true negatives and false negatives, respectively. *sens* is the fraction of positive examples predicted for the interface residue and is given by

$$sens = \frac{TP}{TP + FN}. \qquad (3)$$

spec is the fraction of all positive predictions that are true positives and is given by

$$spec = \frac{TP}{TP + FP}. \quad (4)$$

MCC is given by

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}. \quad (5)$$

Data Sets

We use the dataset from Yan et al.[3] as the benchmark data set which is denoted by YDH77. In YDH77, the sequence lengths of 77 protein chains are more than 10 residues and the sequence identities between these chains are below 30%.[3] 7823 residues are exposed residues in YDH77 and 2352 residues (30.1%) of the exposed residues are in interface. Protein chains in YDH77 are assigned to 6 categories according to the scheme of Chakrabarti et al.[39] The six categories and the number of representative in each category are: protease-inhibitor (11), large protease complexes (6), antigen-antibody (13), enzyme complexes (13), G-protein, cell cycle, signal transduction (16) and miscellaneous (18). In order to check the robust of our method, the dataset used by Liang et al.[6] is also tested. LZZG16 denote this data set because it is composed of 16 protein chains from 8 hetero-dimers.

RESULTS AND DISCUSSION

Comparison results with and without structural homology information

Prediction without structural homology information

Table I shows the performance of our prediction without structurally homology information. Secondary structure information from coding scheme *SS* gives the best performance in *sens*. This result is consistent to the result that binding sites have a preference for β-sheets and for relatively long-structured chains, but not for α-helices.[18] Table I also shows that results with evolutionary information from coding scheme *PSSM* is better than results with amino acid composition information from coding scheme *AA*. Structurally neighboring residue information gives almost the same performance to sequence information when comparing *3D-AA* and *AA*. This result is consistent to the result that the majority of interacting residues are clustered in sequence segments of several contacting residues.[20] The results of coding scheme *3D-PSSM* and of *3D-PSSM'* are similar, it is implied that rescaling profile element values doesn't improve the performance to prediction protein-protein interaction sites.

Prediction with structural homology information

Table II shows results with structural homology information. Evolutionary information from coding scheme *PSSM* gives the best performance. Structural homology information from SCOP contains some evolutionary information[35], so the results that the evolutionary information from coding scheme *PSSM* is with the same performance to the information from coding scheme *AA* when both using structural homology information are reasonable. In Figure 10, the *sens* of 64% proteins are more than 60% and there are 7 proteins with 100% sensitivity.

The performance of structural homology information

When comparing the best results between training scheme D_n and training scheme D_h, D_h is significantly better than D_n. We also have to notice that coding scheme *SS* gives best results with training scheme D_n but coding scheme *PSSM* gives best results with training scheme D_h. *sens* is increased 22%, 24% and 14% by coding scheme *PSSM, AA* and *3D-AA* with structural homology information. But *sens* is decrease 1% by coding scheme *SS* with structural homology information. *spec* is increased 21%, 20%, 24% and 20% with coding scheme *SS*, *PSSM*, *AA* and *3D-AA*. Table I and Table II show that with same coding scheme except *SS*, training scheme D_h is better than training scheme D_n. This suggests that structural homology information is useful to predict protein-protein interaction sites.

Comparison in multiple feature vectors

We try to combine different coding schemes to new feature vector. Table II shows that multiple feature vectors are not helpful to performance. Table II also shows that coding scheme *PSSM* is the best feature vector by itself with training scheme D_h.

Compare with others' works

Table III lists the performance of different method on the benchmark dataset YDH77. Gallet et al.'s method[14] is good on *sens* but not on *spec* while Yan et al.'s two-stage method[3] is good on *spec* but not *sens*. These two method just use amino acid composition information, so we use the results of coding scheme *AA* in order to show the ability of structural homology information. Our method is good not only on *sens* but also on *spec*, so the structural homology information is indeed useful while predict the protein-protein interaction sites. The *MCC* obtained on the class label shuffled dataset (Table III, column 5) is -0.01[3] (as compared with 0.38 on YDH77 by our method) indicating that our method performs significantly better than a random prediction.

Another benchmark dataset

Our method with both evolutionary information and structural homology information is useful on a benchmark YDH77. We use the same training scheme and coding scheme on LZZG16. In table IV, the results are even better when using our method on LZZG16. This shows the robust property of our method.

Cases study

Barstar

The interaction of barnase, an extracellular RNase of Bacillus anylolique-faciens, with its intracellular inhibitor barstar is a good example for protein-protein interaction study because the structures of both the free and the complexed proteins are available at high resolution[17]. When using evolutionary information from with training scheme D_h, the outcome *sens*, *spec* and *accu* of prediction on barstar are 100%, 93.3% and 98.2% respectively. Figure11 and figure 12 show this complex.

G protein β-subunit

1got is a heterotrimeric G protein and the resolution is 2.0 Å. G_β is the β subunit of this G protein[40]. The complex 1got belongs to category "G-proteins, cell cycle, signal transduction"[39]. When the G protein is activated by binding GTP, the heterotrimer dissociates into the α subunit and a βγ heterodimer[29]. Most false negatives or underpredictions of my prediction are in α-βγ interaction region of β subunit, because this region is not interface when G protein is activated. When using evolutionary information from with training scheme D_h, the outcome *sens*, *spec* and *accu* of prediction on barstar are 83%, 78% and 79% respectively.

Others

Six other results from our prediction in different functional categories are showed from figure 13 to figure 24. Figure 13 and figure 14 show barnase. Barnase is from PDB 1brs, chain A. Figure 15 and figure 16 show soybean trypsin inhibitor. Soybean trypsin inhibitor is from PDB 1avw, chain B. Figure 17 and figure 18 show pancreatic trypsin inhibitor (BPTI). BPTI is from PDB 1bth, chain P. Figure 19 and figure 20 show ovomucoid domains from Turkey. This domain is from PDB 1cho, chain I. Figure 21 and figure 22 show VH single domain of an antibody from Human. This domain is from PDB 1ao7, chain B. Figure 23 and 24 show hemagglutinin. Hemagglutinin is from PDB 1qfu, chain A.

CONCOUSIONS

Our structural homology information is beneficial to predict protein-protein interaction sites. Evolutionary information is comparably important by itself when using structural homology training datasets. The sensitivity, specificity, accuracy and Matthew Correlation Coefficient with both evolutionary information and structural homology information on a benchmark dataset YDH77 are 59%, 55%, 74% and 0.38, respectively. Our results are better than those of others for benchmark dataset YDH77.

REFERENCES

1. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44(3):336-343.
2. Li H, Li J. Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. Bioinformatics 2005;21(3):314-324.
3. Yan C, Dobbs D, Honavar V. A two-stage classifier for identification of protein-protein interface residues. Bioinformatics 2004;20 Suppl 1:I371-I378.
4. Fernandez A, Scheraga HA. Insufficiently dehydrated hydrogen bonds as

determinants of protein interactions. Proc Natl Acad Sci U S A 2003;100(1):113-118.

5. Josephson K, Logsdon NJ, Walter MR. Crystal structure of the IL-10/IL-10R1 complex reveals a shared receptor binding site. Immunity 2001;15(1):35-46.

6. Liang S, Zhang J, Zhang S, Guo H. Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores. Proteins 2004;57(3):548-557.

7. Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science 1985;228(4705):1315-1317.

8. Rodi DJ, Agoston GE, Manon R, Lapcevich R, Green SJ, Makowski L. Identification of small molecule binding sites within proteins using phage display technology. Comb Chem High Throughput Screen 2001;4(7):553-572.

9. Sidhu SS, Fairbrother WJ, Deshayes K. Exploring protein-protein interactions with phage display. Chembiochem 2003;4(1):14-25.

10. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001;17(3):284-285.

11. Botstein D, Shortle D. Strategies and applications of in vitro mutagenesis. Science 1985;229(4719):1193-1201.

12. Clemmons DR. Use of mutagenesis to probe IGF-binding protein structure/function relationships. Endocr Rev 2001;22(6):800-817.

13. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272(1):133-143.

14. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol 2000;302(4):917-926.

15. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein--protein

interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269(5):1356-1361.

16. Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 1998;26(1):313-315.

17. Gao Y, Wang R, Lai L. Structure-based method for analyzing protein-protein interfaces. J Mol Model (Online) 2004;10(1):44-54.

18. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338(1):181-199.

19. Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel 2004;17(2):165-173.

20. Ofran Y, Rost B. Predicted protein-protein interaction sites from local sequence information. FEBS Lett 2003;544(1-3):236-239.

21. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 2003;326(1):255-261.

22. Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.

23. Huang SW, Hwang JK. Computation of conformational entropy from protein sequences using the machine-learning method--application to the study of the relationship between structural conservation and local structural stability. Proteins 2005;59(4):802-809.

24. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 2004;13(5):1402-1406.

25. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M,

Jr., Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 2000;97(1):262-267.

26. Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng 2003;16(8):553-560.

27. Chen YC, Lin YS, Lin CJ, Hwang JK. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins 2004;55(4):1036-1042.

28. Yu CS, Wang JY, Yang JM, Lyu PC, Lin CJ, Hwang JK. Fine-grained protein fold assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets. Proteins 2003;50(4):531-536.

29. Branden C, Tooze J. Introduction to Protein Structure; 1999. 252 p.

30. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2.8 ed; 2001. p LIBSVM: A library for support vector machines.

31. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20(3):216-226.

32. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, Jones DT. Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 2004;342(1):307-320.

33. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.

34. Hubbard TJ, Murzin AG, Brenner SE, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 1997;25(1):236-239.

35. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a Structural

Classification of Proteins database. Nucleic Acids Res 1999;27(1):254-256.

36. Myers EW, Miller W. Optimal alignments in linear space. Comput Appl Biosci 1988;4(1):11-17.

37. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 2001;29(1):228-229.

38. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405(2):442-451.

39. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47(3):334-343.

40. Lambright DG, Sondek J, Bohm A, Skiba NP, Hamm HE, Sigler PB. The 2.0 A crystal structure of a heterotrimeric G protein. Nature 1996;379(6563):311-319.

TABLE I: Performance without structural homology information

	SS	*RSA*	*PSSM*	*AA*	*3D-PSSM*	*3D-PSSM'*	*3D-AA*
sens	0.50	0.42	0.37	0.33	0.32	0.32	0.31
spec	0.33	0.32	0.35	0.32	0.33	0.32	0.32
accu	0.54	0.58	0.60	0.59	0.60	0.60	0.59
MCC	0.06	0.04	0.06	0.03	0.04	0.02	0.02

On testing dataset YDH77 and results are sorted by *sens*.

SS: information from protein secondary structure

PSSM: evolutionary information

AA: amino acid composition information

3D-AA: information from protein tertiary structure

3D-PSSM: evolutionary information from structurally neighboring residues

3D-PSSM': evolutionary information from structurally neighboring residues

TABLE II: **Performance with structural homology training datasets**

	PSSM	*AA*	*RSA*	*SS*	*3D-AA*	*AA+RSA*	*AA+SS*	*AA+PSSM*	*SS+PSSM*	*AA+SS*	*AA+SS*
										+PSSM	*+PSSM+RSA*
sens	0.59	0.57	0.58	0.49	0.45	0.57	0.57	0.54	0.56	0.55	0.54
spec	0.55	0.56	0.45	0.56	0.52	0.55	0.55	0.54	0.56	0.56	0.57
accu	0.74	0.74	0.66	0.68	0.72	0.74	0.74	0.74	0.75	0.75	0.75
MCC	0.38	0.37	0.26	0.27	0.29	0.38	0.37	0.36	0.38	0.38	0.38

On testing dataset YDH77 and results are sorted by *sens*.

PSSM: evolutionary information

AA: amino acid composition information

RSA: information from relative solvent accessibility

SS: information from secondary structure

3D-AA: information from protein tertiary structure

TABLE III: Performances of different methods

	Coding scheme *AA* with training scheme D_h	Method of Gallet et al.[a]	SVM Method of Yan et al.	Two-stage method of Yan et al.	Two-stage method[b] of Yan et al.
sens	0.57	0.44	0.43	0.39	0.37
spec	0.56	0.30	0.44	0.58	0.31
accu	0.74	0.51	0.66	0.72	0.53
MCC	0.37	-0.02	0.19	0.30	-0.01

On testing dataset YDH77 and results are sorted by *sens*.

[a] Data from Yan et al. using method of Gallet et al.

[b] Class label are randomly shuttled for all the examples before training and testing the classifier

TABLE IV: Performances of different testing datasets

	YDH77	LZZG16
sens	0.57	0.62
spec	0.56	0.68
accu	0.75	0.79
MCC	0.38	0.49

Using coding scheme *PSSM* and training scheme D_h

TABLE V(a): 11 protease-inhibitors

Protein chain	Description	*sens*	*spec*	*accu*	*MCC*
1hiaAB	Hirustasin	43.23%	90.91%	26.09%	0.20
d1avwb_	Soybean Trypsin Inhibitor, Orthorhomic Crystal Form	81.40%	45.45%	71.43%	0.46
d1choi_	Alpha-chymotrypsin Complex with OMTKY3	88.89%	81.82%	81.82%	0.74
d1flei_	Elafin	45.16%	0.00%	0.00%	0.00
d1hiai_	Hirustasin	94.59%	100.00%	86.67%	0.89
d1stfe_	Papain (E.C. 3.4.22.2)	82.29%	63.64%	35.00%	0.38
d1stfi_	Papain (E.C. 3.4.22.2)	85.45%	68.18%	93.75%	0.70
d1tgsi_	Tripsinogen Complex with Trypsin Inhibitor	77.50%	75.00%	70.59%	0.54
d2sici_	Subtilisin (serine protease)	100.00%	100.00%	100.00%	1.00
d3sgbe_	Proteinase B from Streptomyces Criseus(SGPB)	87.80%	58.82%	76.92%	0.60
d4cpai_	Carboxypeptidase A (Cox) Complexe with Cox Inhibitor	64.00%	50.00%	11.11%	0.09

TABLE V(b): 6 large protease complexes

Protein chain	Description	*sens*	*spec*	*accu*	*MCC*
1danTU	Soluble Tissue Factor	46.03%	66.67%	21.05%	0.06
d1bthp_	Bovine Pancreatic Trpsin Inhibitor	83.72%	100.00%	65.00%	0.71
d1tbqr_	Rhodniin	48.05%	47.37%	23.08%	-0.04
d1tocb_	Thrombin	48.98%	75.00%	16.07%	0.13
d1tocr_	Ornithodorin	57.47%	55.56%	37.50%	0.13
d4htci_	Thrombin Complex With Recombinant Hirudin	94.00%	93.94%	96.88%	0.87

TABLE IV(c): 13 enzyme complexes

Protein chain	Description	*sens*	*spec*	*accu*	*MCC*
d1brsa_	Barnase (G specific endonuclease)	50.82%	44.44%	61.54%	0.04
d1brsd_	Barstar	98.15%	93.33%	100.00%	0.95
d1dfje_	Ribonuclease A	43.48%	33.33%	34.48%	-0.15
d1dfji_	Ribonuclease Inhibitor	72.86%	27.42%	58.62%	0.26
d1dhka_	Procine Pancreatic - Amylase	82.83%	33.33%	6.25%	0.08
d1dhkb_	Bean Lectin-like Inhibitor	50.00%	18.75%	22.22%	-0.16
d1fssa_	Acetycholinesterase (E.C. 3.1.1.7)	89.71%	0.00%	0.00%	0.00
d1fssb_	Fasciculin II	95.45%	94.12%	94.12%	0.90
d1glaf_	Glycerol Kinase	95.29%	80.00%	100.00%	0.87
d1glag_	Glycerol Kinase	77.38%	30.77%	12.12%	0.08
d1udie_	Uracil-DNA Glycosylase	85.19%	100.00%	23.81%	0.45
d1udii_	Uracil-DNA Glycosylase Inhibitor	90.91%	86.36%	90.48%	0.81
d1ydre_	C-Amp-Dependent Protein Kinase	95.89%	76.00%	100.00%	0.85

TABLE IV(d): 13 antigen-antibody

Protein chain	Description	*sens*	*spec*	*accu*	*MCC*
1kb5AB	Kb5-C20 T-Cell Antigen Receptor (antibody)	69.12%	62.22%	52.83%	0.34
2jelLH	Jel42 Fab Fragment (antibody)	70.75%	54.55%	51.85%	0.32
d1ao7a_	Hia-A 0201	80.65%	71.43%	68.63%	0.56
d1ao7b_	Beta-2 Microglobulin	90.77%	84.38%	96.43%	0.82
d1jhla_	Fv Fragment (antibody)	61.19%	12.50%	14.29%	-0.12
d1mela_	Vh Single-Domain Antiboy	83.10%	56.25%	64.29%	0.50
d1nfdb_	T-Cell Receptor(antigen)	55.97%	60.78%	44.29%	0.13
d1nmbn_	N9 Neuraminidase	82.58%	34.15%	100.00%	0.53
d1nsns_	Staphylococcal Nuclease	67.57%	0.00%	0.00%	-0.11
d1ospo_	Outer Surface Protein A	71.53%	0.00%	0.00%	-0.16
d1qfua_	Hemagglutinin (antigen)	69.05%	63.41%	70.27%	0.38
d1qfub_	Hemagglutinin	65.57%	57.89%	96.49%	0.42
d2jelp_	Histindine-Contating Protein (Jel42 Fab/Hpr complex)	23.53%	4.17%	5.88%	-0.58

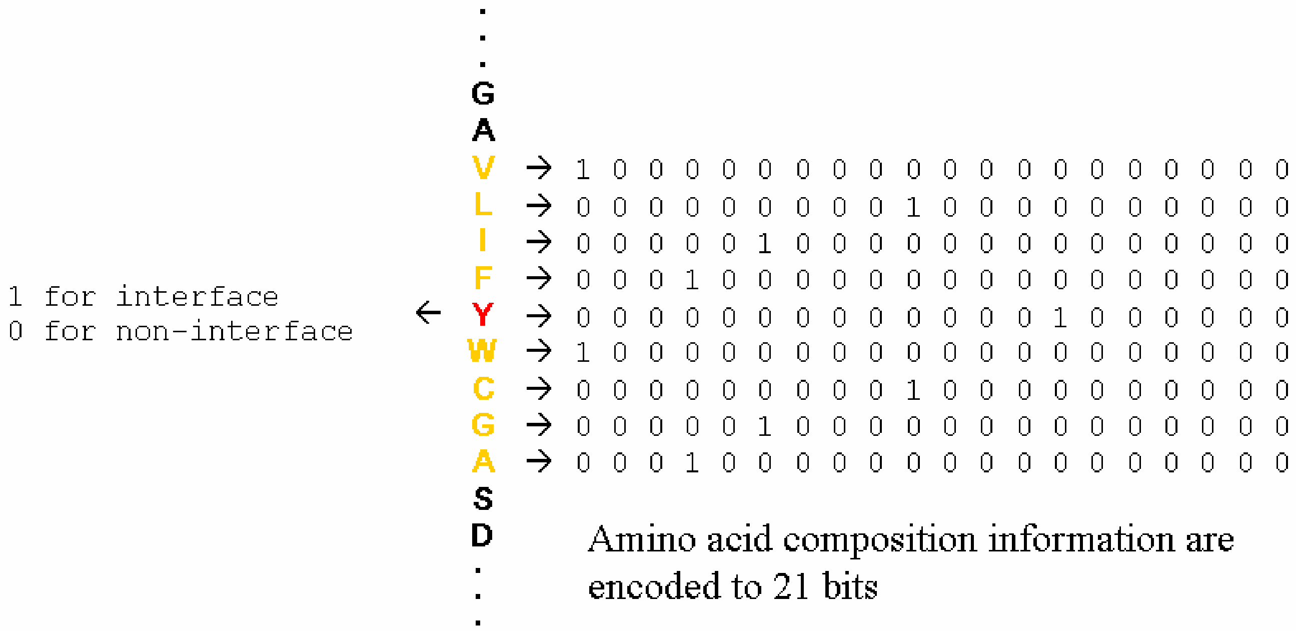
TABLE IV(e): 16 G-protein, cell cycle, signal transduction

Protein chain	Description	*sens*	*spec*	*accu*	*MCC*
d1a0oa_	CheY	94.44%	81.25%	92.86%	0.83
d1a0ob_	CheA	97.87%	100.00%	93.33%	0.95
d1a2ka_	Nuclear Transport Factor 2	78.67%	85.71%	73.17%	0.58
d1a2kc_	Ras-family GTPase Ran	64.58%	21.21%	46.67%	0.11
d1agra_	Guanine Nucleotide-Binding Protein G(I)	71.60%	23.53%	26.67%	0.08
d1agre_	Rgs4(regulator of guanine nucleotide-binding protein)	86.49%	76.92%	58.82%	0.59
d1aipa_	Elogation Factor Tu	80.56%	0.00%	0.00%	-0.07
d1aipc_	Elogation Factor Ts	64.17%	76.67%	38.98%	0.32
d1fina_	Cycline-dependent Kinase 2	91.43%	86.67%	86.67%	0.80
d1finb_	Cycline A	57.26%	31.71%	37.14%	0.03
d1gotb_	β subunit of G protein	79.35%	76.83%	82.89%	0.59
d1guaa_	Rap1A (one member of Ras family)	78.16%	36.84%	50.00%	0.30
d1guab_	C-Raf1	93.48%	82.35%	100.00%	0.86
d1tx4a_	P50-Rhogap(GTPase-activating protein rhogap)	88.12%	65.52%	90.48%	0.70
d1tx4b_	Transforming Protein Rhoa	71.59%	51.85%	53.85%	0.33
d2trcp_	Phosducin	79.71%	72.22%	59.09%	0.51

TABLE IV(f): 13 miscellaneous

Protein chain	Description	*sens*	*spec*	*accu*	*MCC*
1sebAB	Hia Class II Histocompatibility Antigen	52.81%	67.37%	45.07%	0.10
d1ak4a_	Cyclophilin A	69.86%	27.59%	88.89%	0.38
d1ak4c_	HIV-1 Capsid	94.32%	91.67%	88.00%	0.86
d1atna_	Deoxyribonuclease I Complex with Actin	57.89%	19.05%	80.00%	0.22
d1atnd_	Deoxyribonuclease I Complex with Actin	84.40%	0.00%	0.00%	0.00
d1dkga_	Nucleotide Exchange Factor Grpe	68.91%	78.57%	63.77%	0.39
d1dkgd_	Molecular Chaperone Dnak	82.32%	18.18%	8.00%	0.03
d1efna_	Fyn Tryosine Kinase	91.67%	78.57%	100.00%	0.83
d1efnb_	HIV-1 Nef Protein (SH3 domain)	93.33%	86.96%	95.24%	0.86
d1fc2c_	Immunoglobulin Fc and Fragment B	75.76%	61.11%	91.67%	0.56
d1fc2d_	Immunoglobin Fc and Fragment B (antibody)	63.64%	11.63%	45.45%	0.07
d1hwga_	Human Growth Hormone	92.93%	90.24%	92.50%	0.85
d1hwgb_	Growth Hormone Binding Protein	67.62%	55.56%	14.29%	0.14

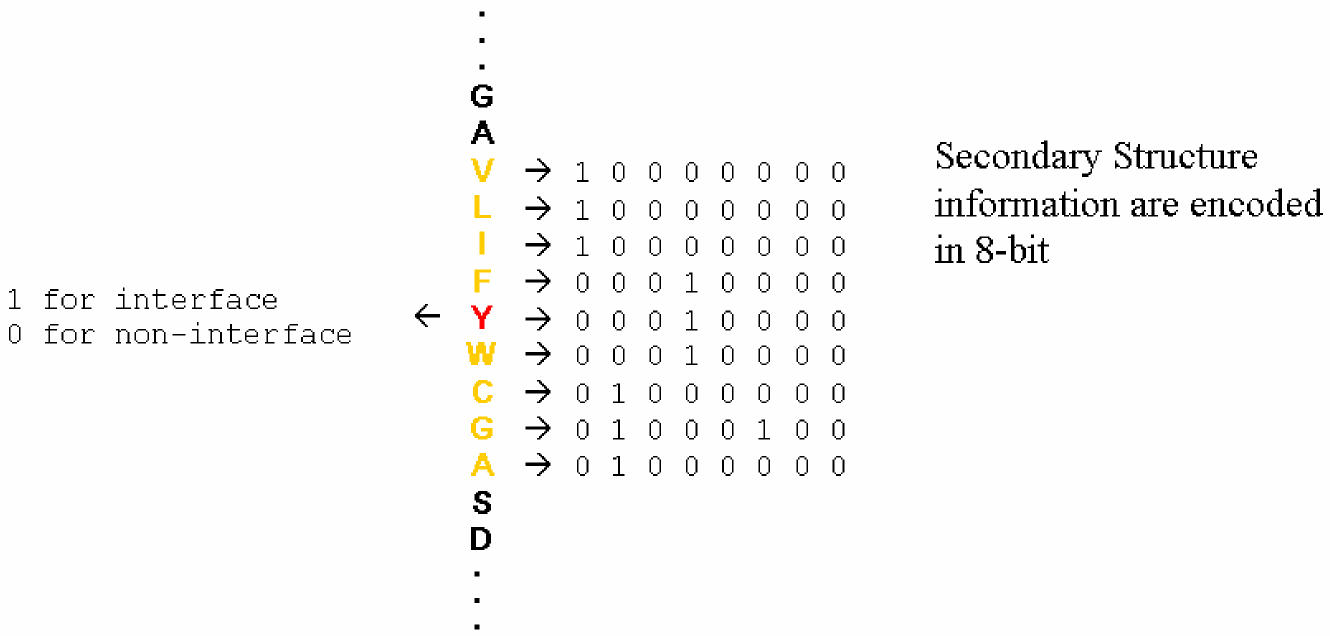
Figure 1



```
                             .
                             .
                             .
                             G
                             A
                             V → 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                             L → 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
                             I → 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 for interface              F → 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 for non-interface        ← Y → 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
                             W → 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                             C → 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
                             G → 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                             A → 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                             S
                             D
                             .
                             .
                             .
```

Amino acid composition information are encoded to 21 bits

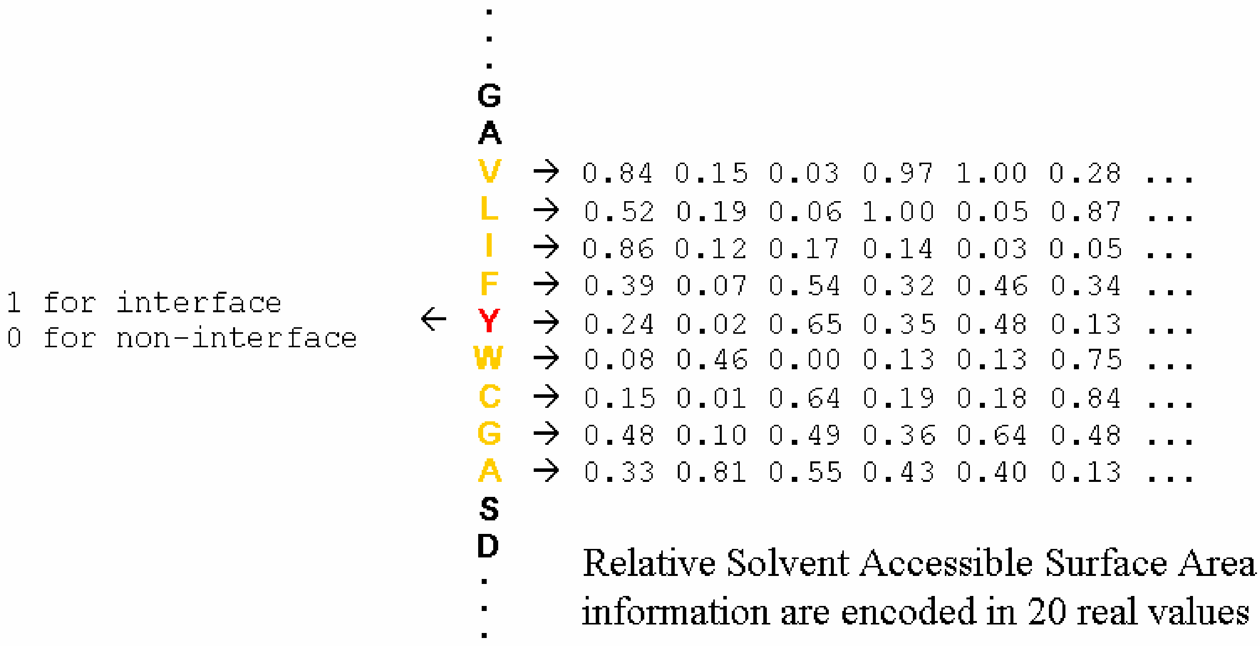
Figure 2



```
V → 1 0 0 0 0 0 0 0
L → 1 0 0 0 0 0 0 0
I → 1 0 0 0 0 0 0 0
F → 0 0 0 1 0 0 0 0
Y → 0 0 0 1 0 0 0 0
W → 0 0 0 1 0 0 0 0
C → 0 1 0 0 0 0 0 0
G → 0 1 0 0 0 1 0 0
A → 0 1 0 0 0 0 0 0
```

Secondary Structure information are encoded in 8-bit

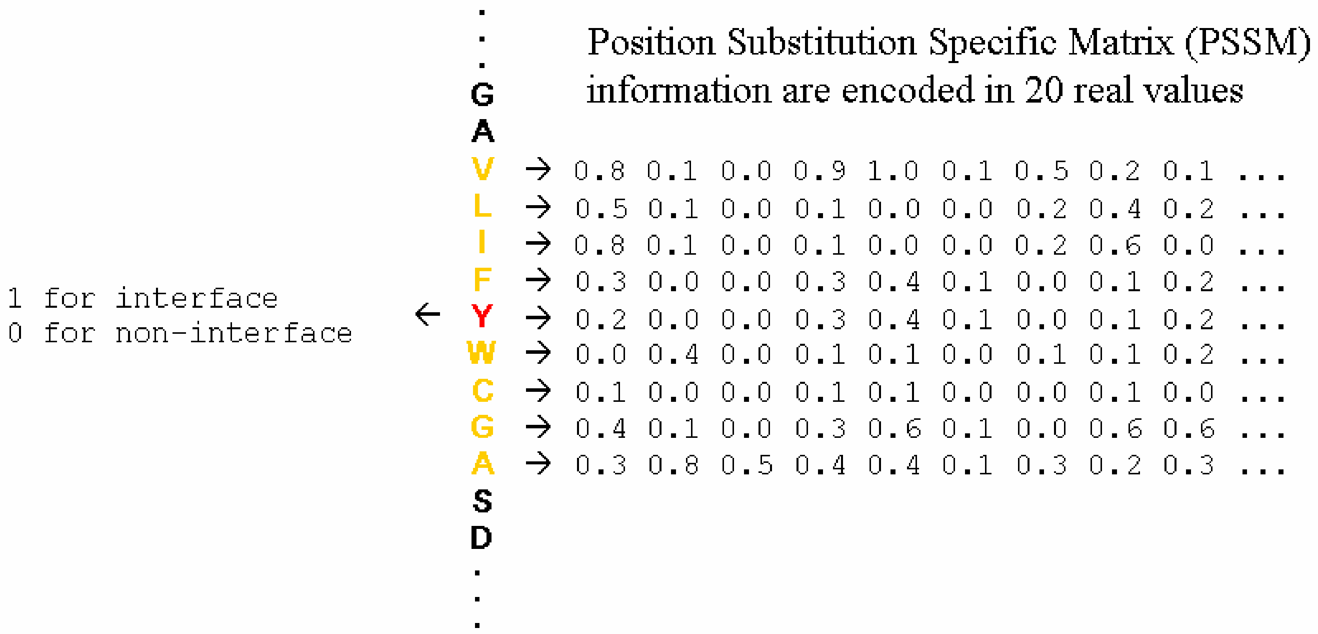
Figure 3



```
V → 0.84 0.15 0.03 0.97 1.00 0.28 ...
L → 0.52 0.19 0.06 1.00 0.05 0.87 ...
I → 0.86 0.12 0.17 0.14 0.03 0.05 ...
F → 0.39 0.07 0.54 0.32 0.46 0.34 ...
Y → 0.24 0.02 0.65 0.35 0.48 0.13 ...
W → 0.08 0.46 0.00 0.13 0.13 0.75 ...
C → 0.15 0.01 0.64 0.19 0.18 0.84 ...
G → 0.48 0.10 0.49 0.36 0.64 0.48 ...
A → 0.33 0.81 0.55 0.43 0.40 0.13 ...
```

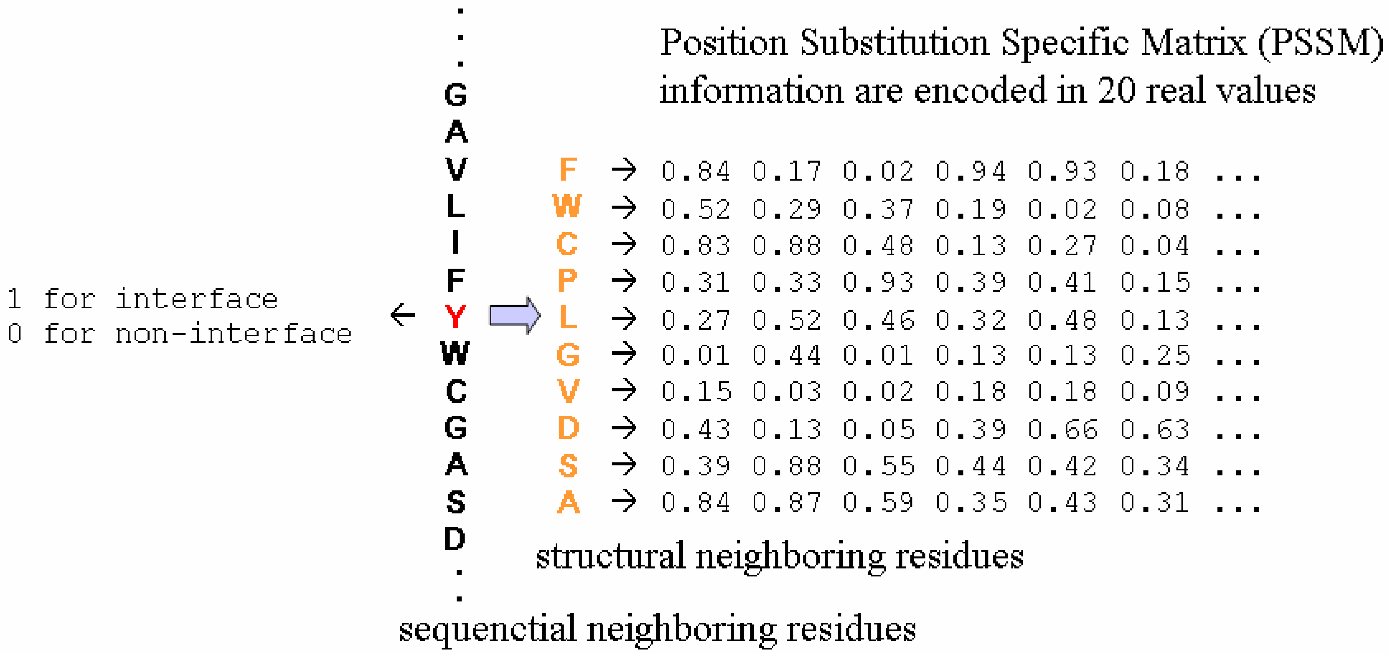
Relative Solvent Accessible Surface Area
information are encoded in 20 real values

Figure 4


Position Substitution Specific Matrix (PSSM)
information are encoded in 20 real values
1 for interface
0 for non-interface
←
.
.
.
G
A
V
L
I
F
Y
W
C
G
A
S
D
.
.
.

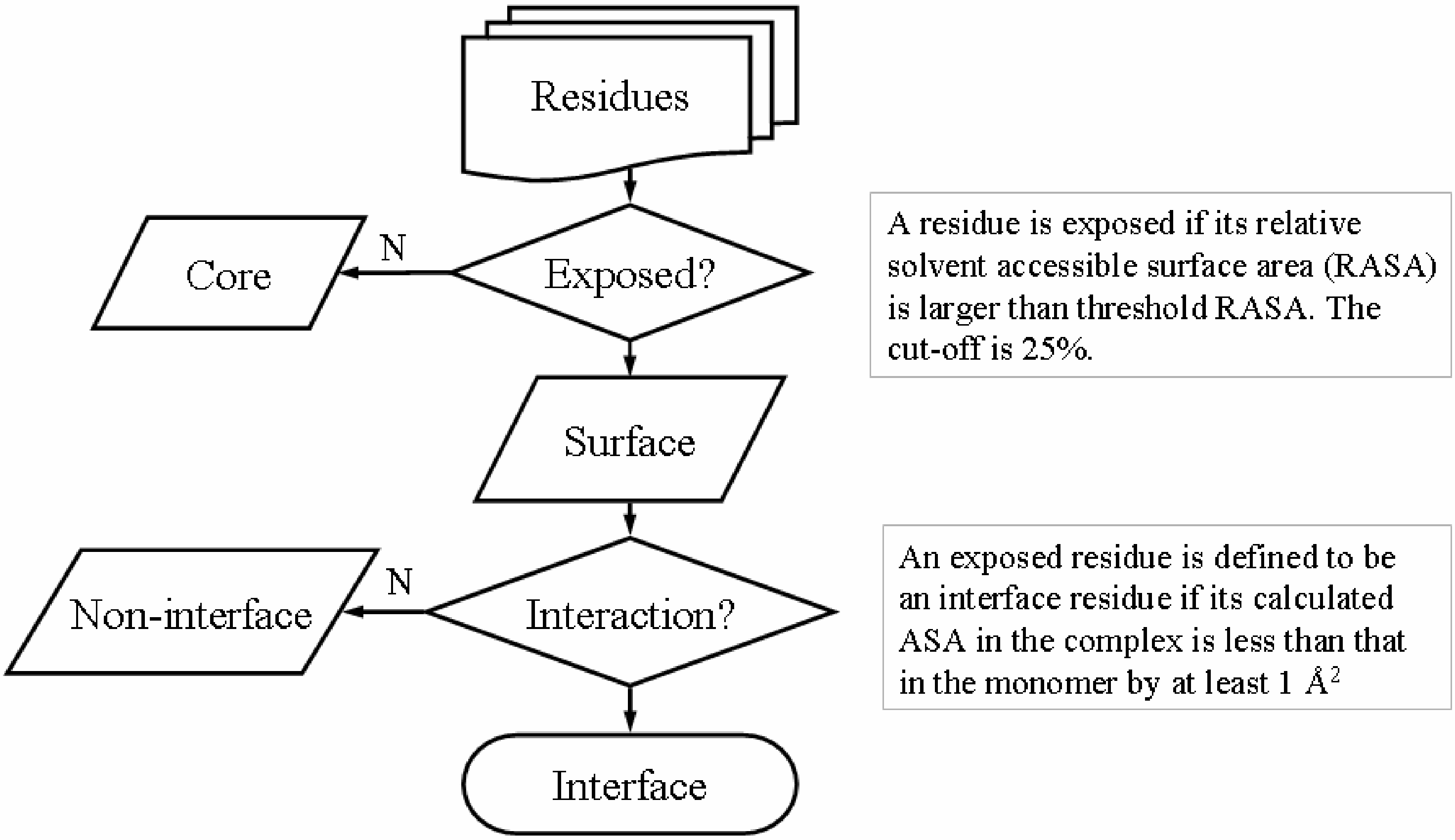
```
V → 0.8 0.1 0.0 0.9 1.0 0.1 0.5 0.2 0.1 ...
L → 0.5 0.1 0.0 0.1 0.0 0.0 0.2 0.4 0.2 ...
I → 0.8 0.1 0.0 0.1 0.0 0.0 0.2 0.6 0.0 ...
F → 0.3 0.0 0.0 0.3 0.4 0.1 0.0 0.1 0.2 ...
Y → 0.2 0.0 0.0 0.3 0.4 0.1 0.0 0.1 0.2 ...
W → 0.0 0.4 0.0 0.1 0.1 0.0 0.1 0.1 0.2 ...
C → 0.1 0.0 0.0 0.1 0.1 0.0 0.0 0.1 0.0 ...
G → 0.4 0.1 0.0 0.3 0.6 0.1 0.0 0.6 0.6 ...
A → 0.3 0.8 0.5 0.4 0.4 0.1 0.3 0.2 0.3 ...
```

Figure 5


1 for interface
0 for non-interface
.
.
.
G
A
V
L
I
F
Y
W
C
G
A
S
D
.
.
Position Substitution Specific Matrix (PSSM)
information are encoded in 20 real values
structural neighboring residues
sequenctial neighboring residues

F → 0.84 0.17 0.02 0.94 0.93 0.18 ...
W → 0.52 0.29 0.37 0.19 0.02 0.08 ...
C → 0.83 0.88 0.48 0.13 0.27 0.04 ...
P → 0.31 0.33 0.93 0.39 0.41 0.15 ...
L → 0.27 0.52 0.46 0.32 0.48 0.13 ...
G → 0.01 0.44 0.01 0.13 0.13 0.25 ...
V → 0.15 0.03 0.02 0.18 0.18 0.09 ...
D → 0.43 0.13 0.05 0.39 0.66 0.63 ...
S → 0.39 0.88 0.55 0.44 0.42 0.34 ...
A → 0.84 0.87 0.59 0.35 0.43 0.31 ...

Figure 6



A residue is exposed if its relative solvent accessible surface area (RASA) is larger than threshold RASA. The cut-off is 25%.

An exposed residue is defined to be an interface residue if its calculated ASA in the complex is less than that in the monomer by at least 1 Å^2

Figure 7

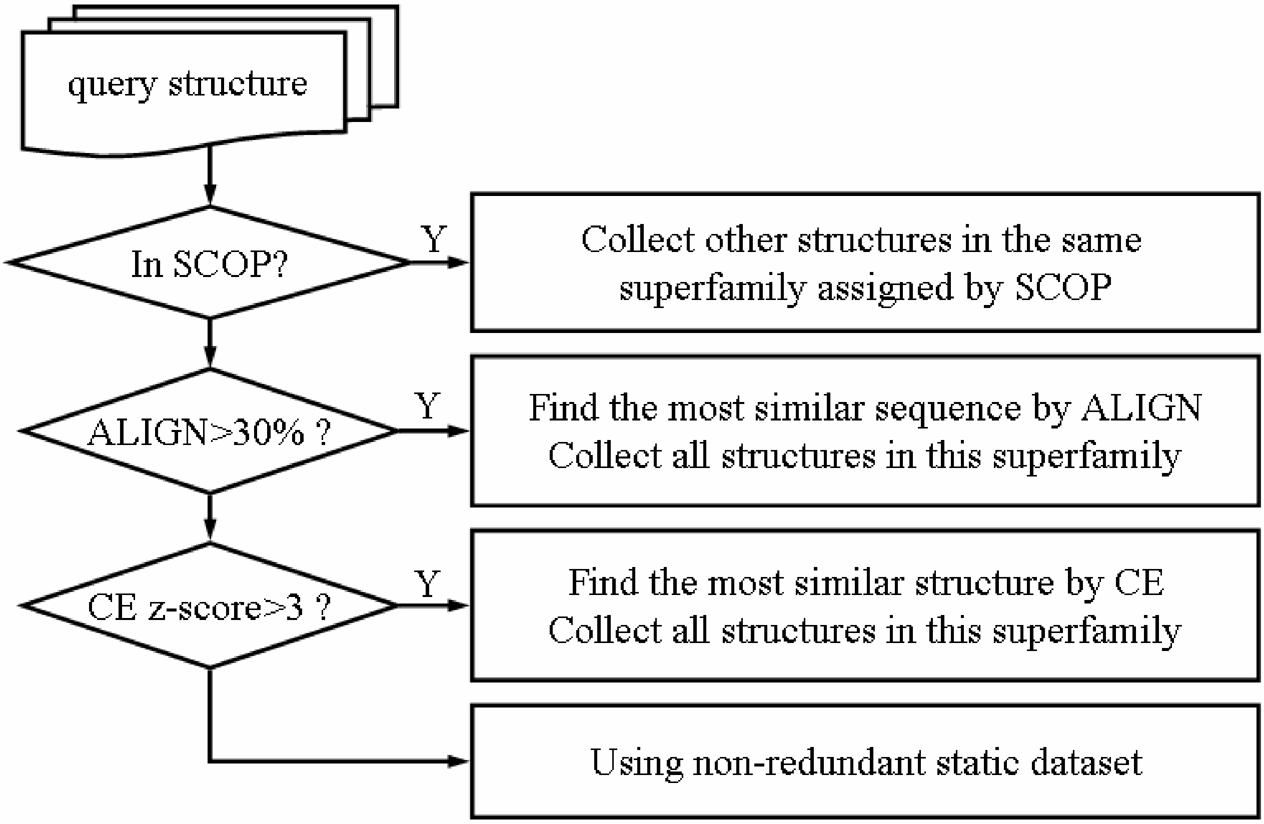

query structure
In SCOP?
Y
Collect other structures in the same superfamily assigned by SCOP
ALIGN>30% ?
Y
Find the most similar sequence by ALIGN
Collect all structures in this superfamily
CE z-score>3 ?
Y
Find the most similar structure by CE
Collect all structures in this superfamily
Using non-redundant static dataset

Figure 8

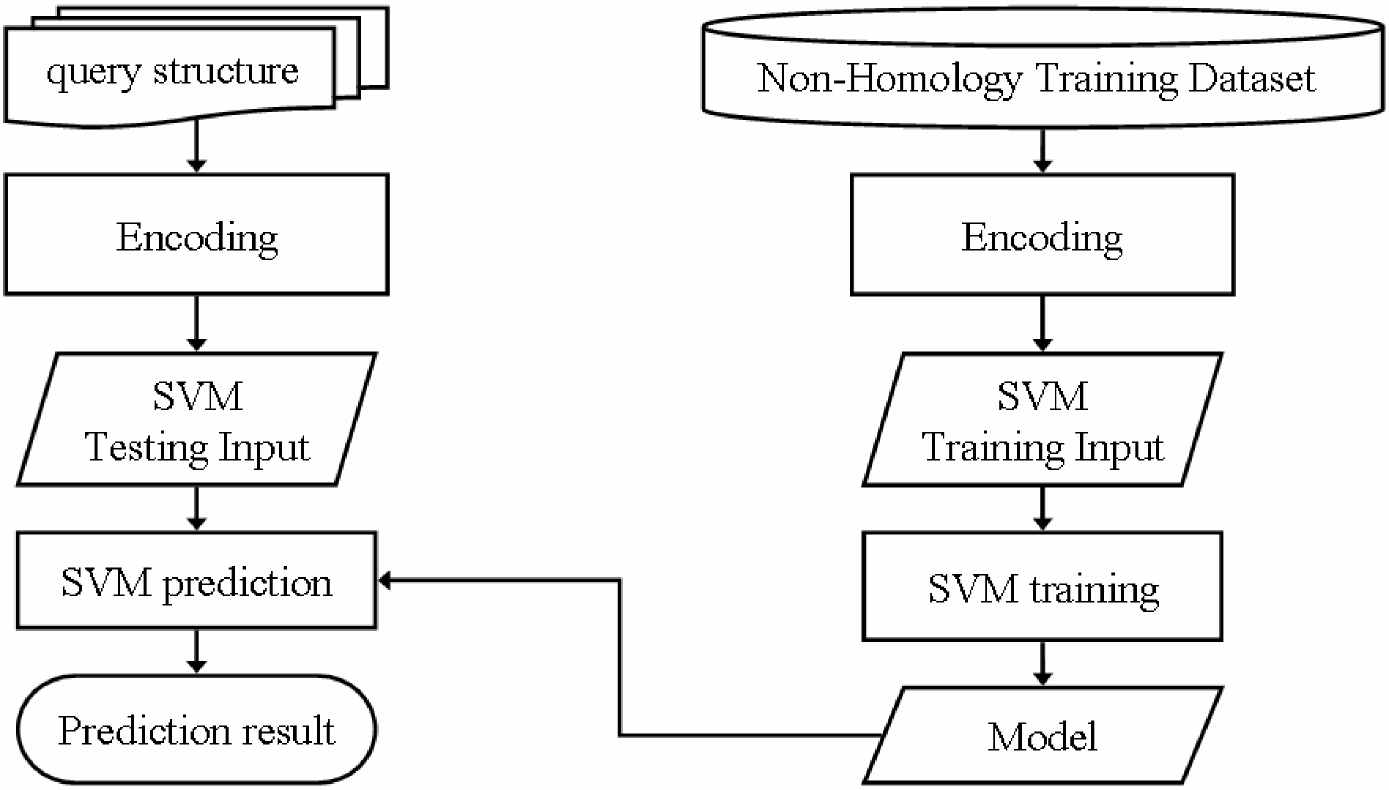

query structure
Encoding
SVM
Testing Input
SVM prediction
Prediction result
Non-Homology Training Dataset
Encoding
SVM
Training Input
SVM training
Model

Figure 9

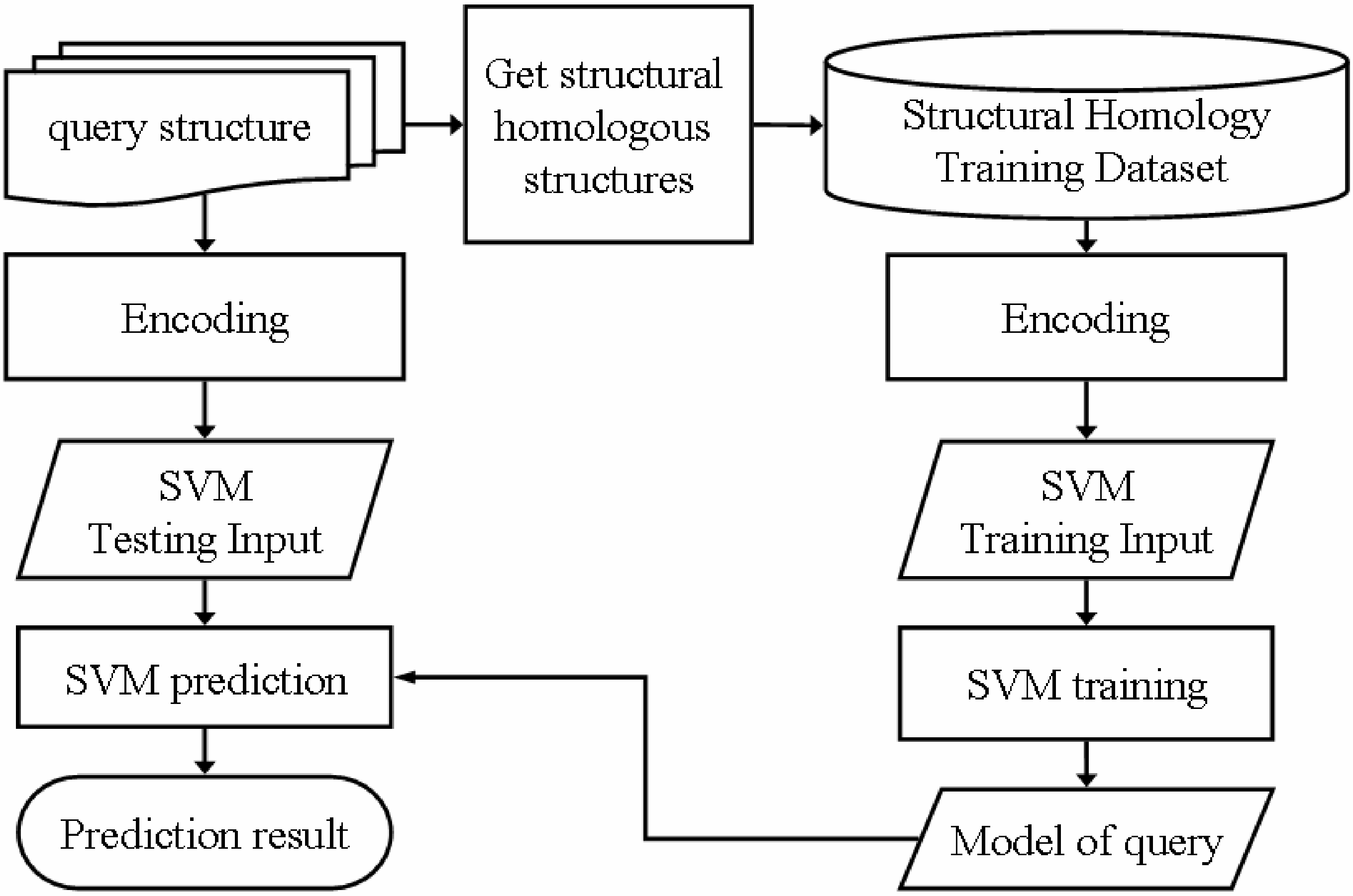

query structure
Get structural homologous structures
Structural Homology Training Dataset
Encoding
Encoding
SVM Testing Input
SVM Training Input
SVM prediction
SVM training
Prediction result
Model of query

Figure 10

Prediction by coding scheme *PSSM* with structural homology information

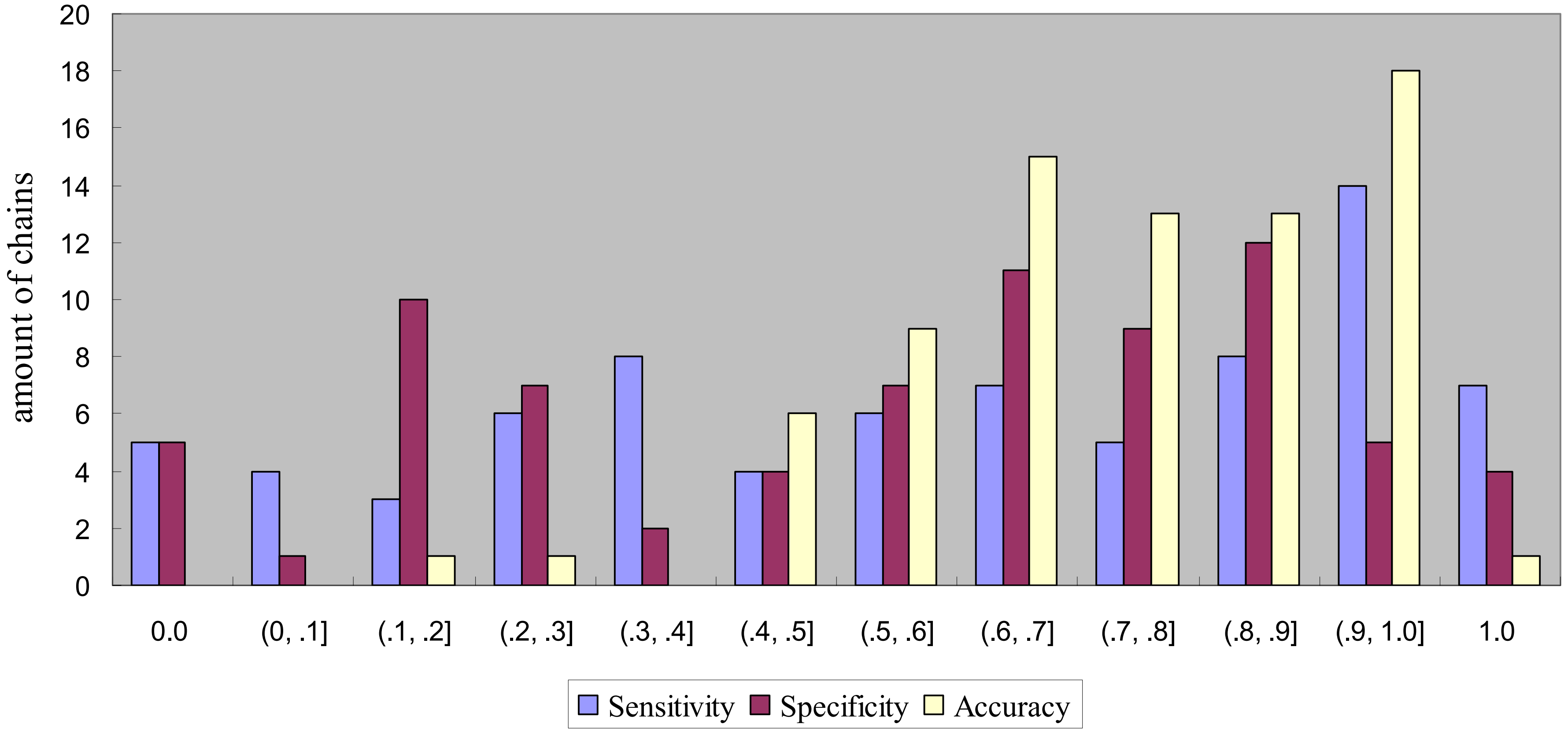


Figure 11

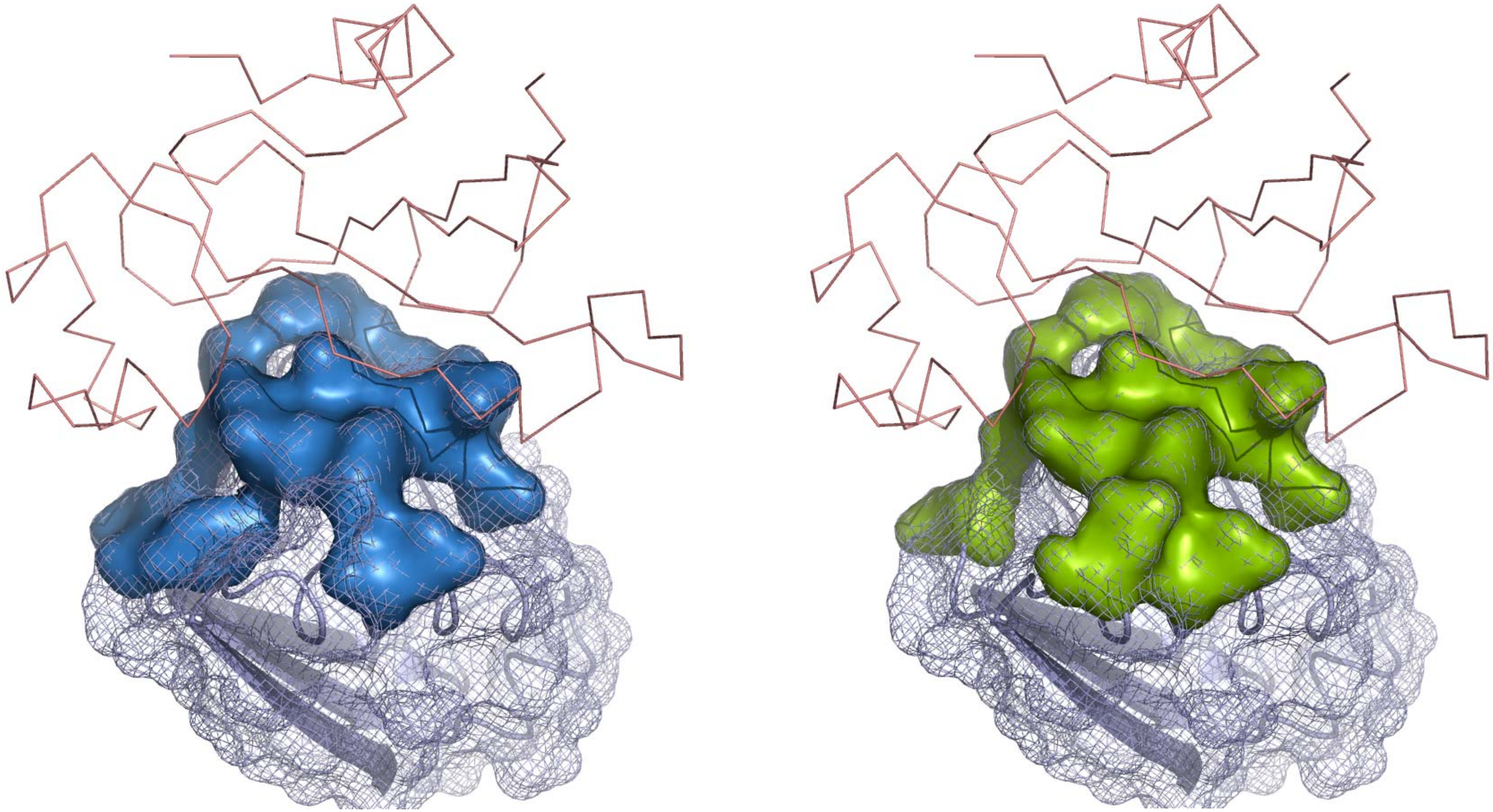


Figure 12

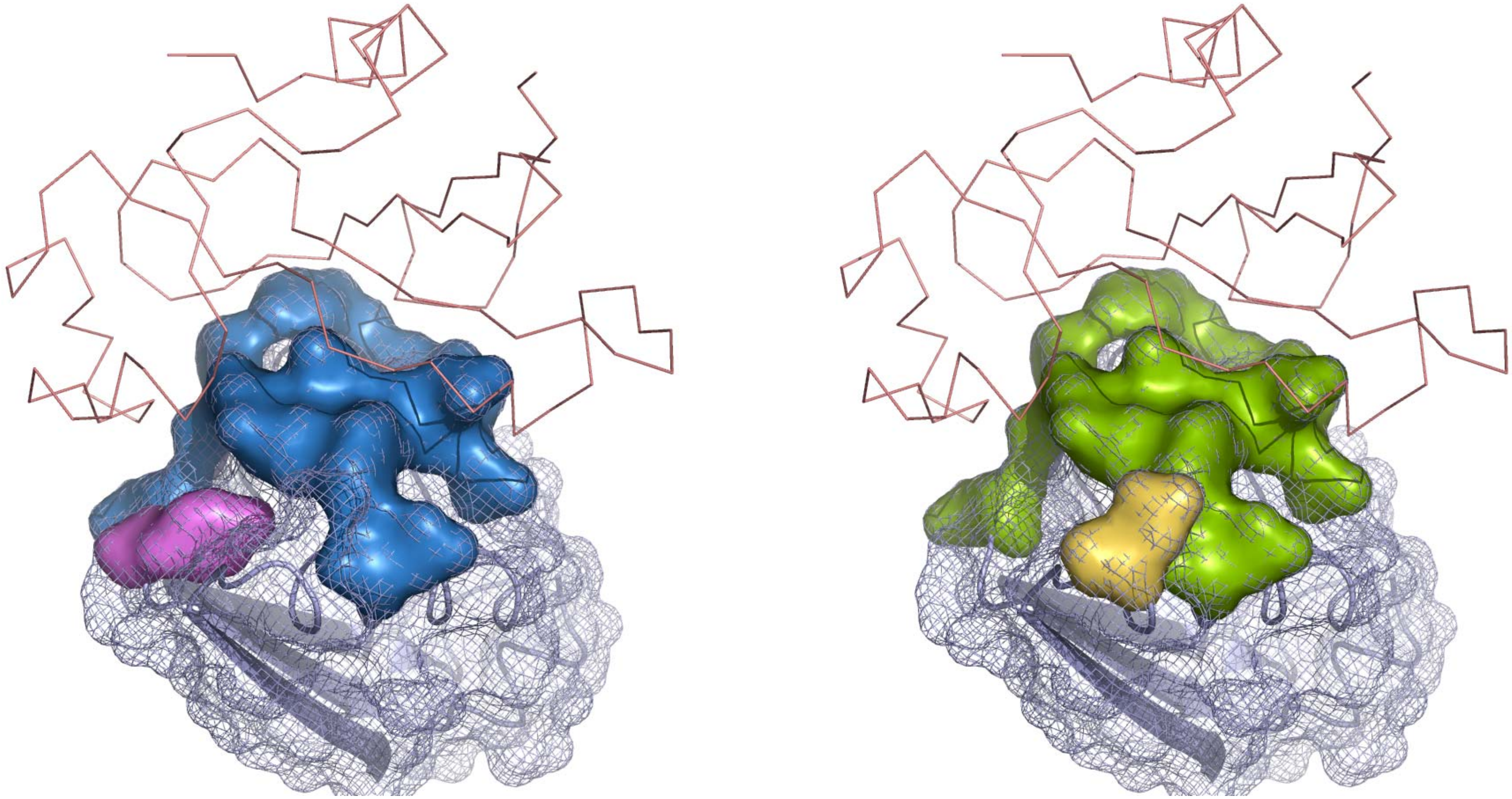


Figure 13

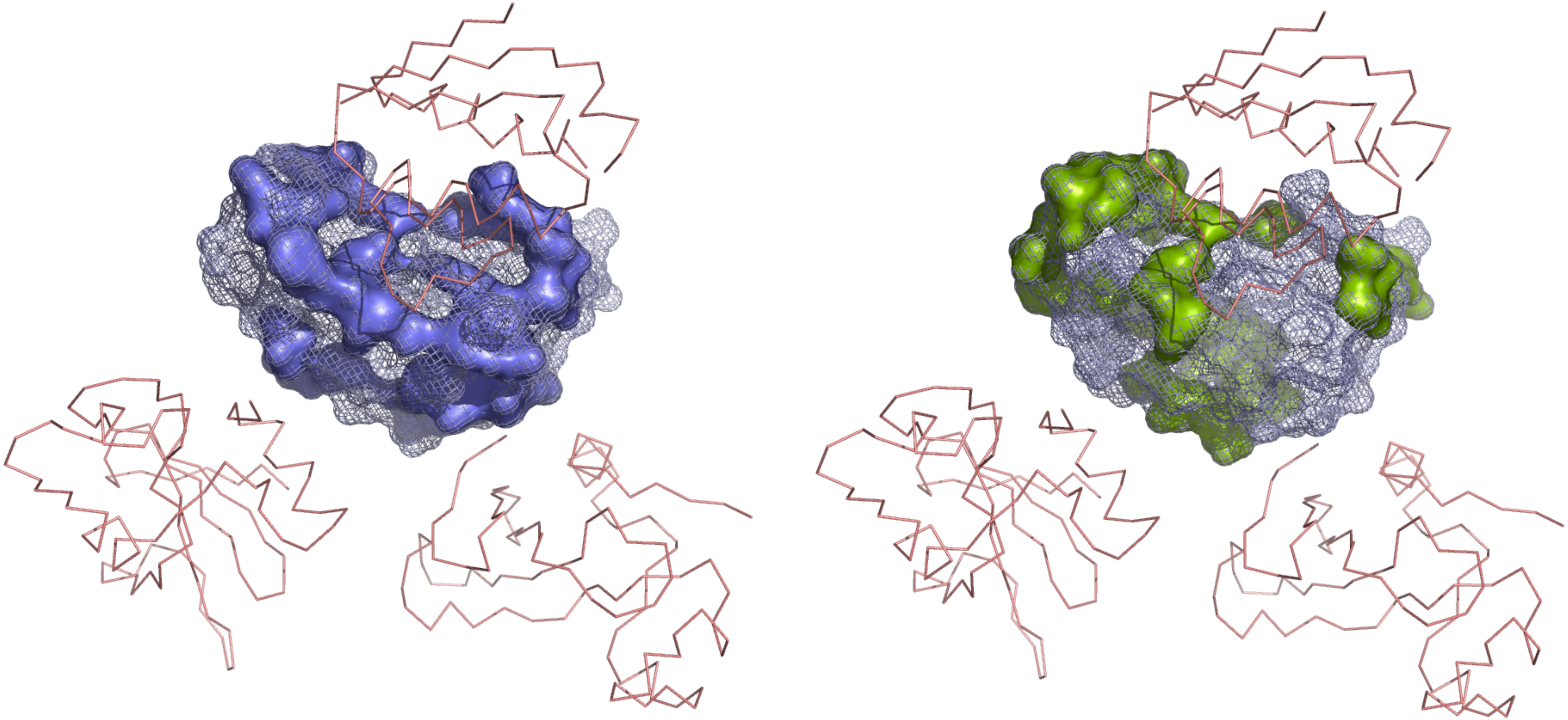


Figure 14

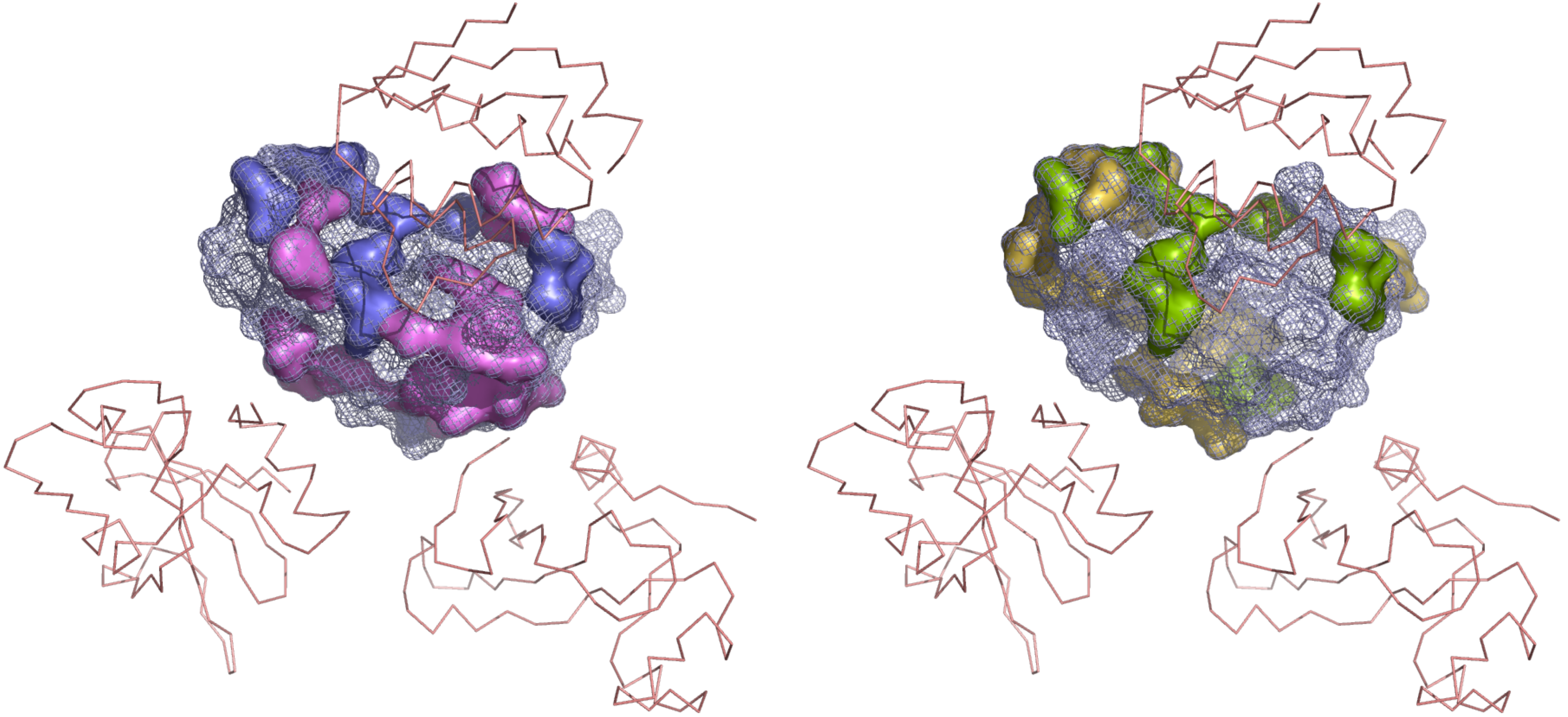


Figure 15

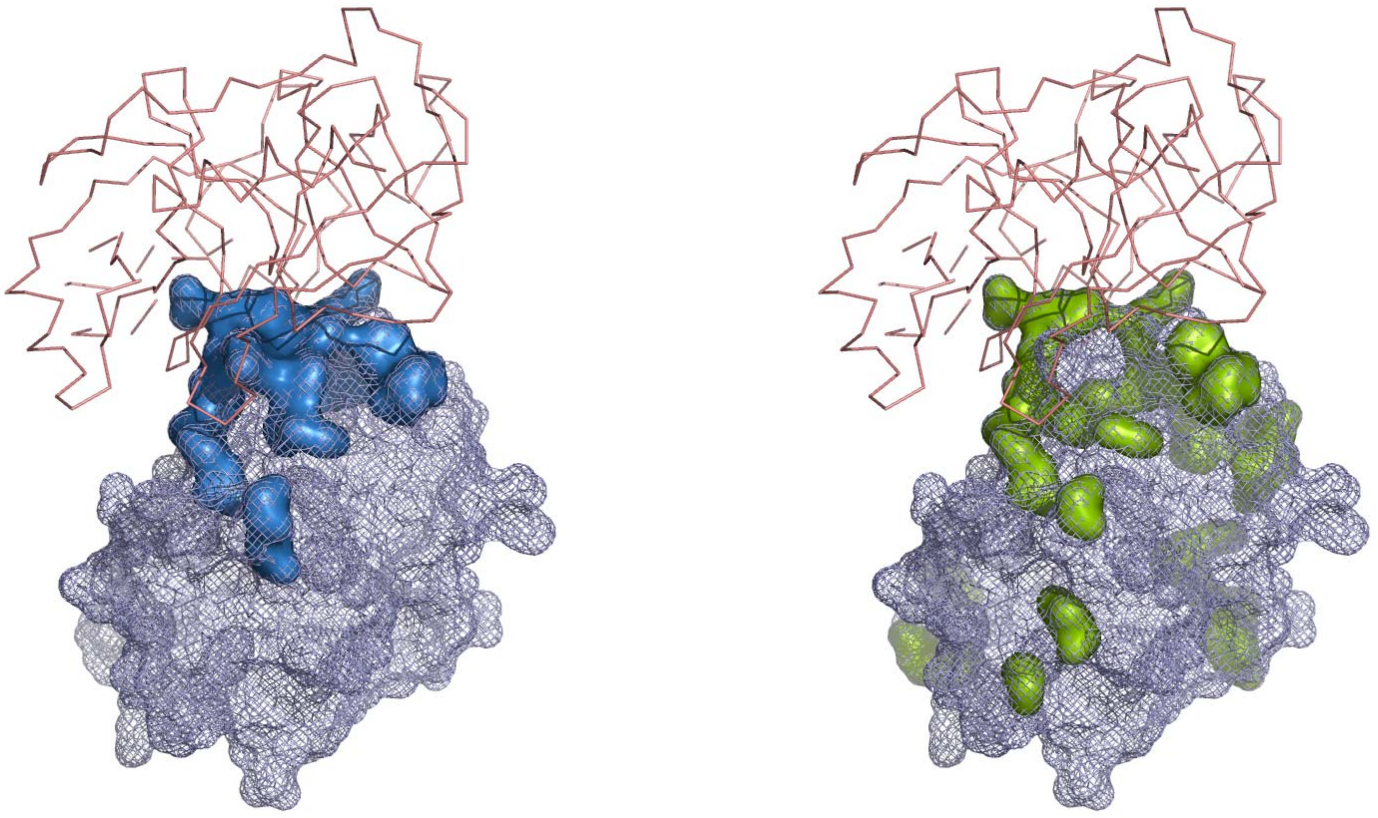


Figure 16

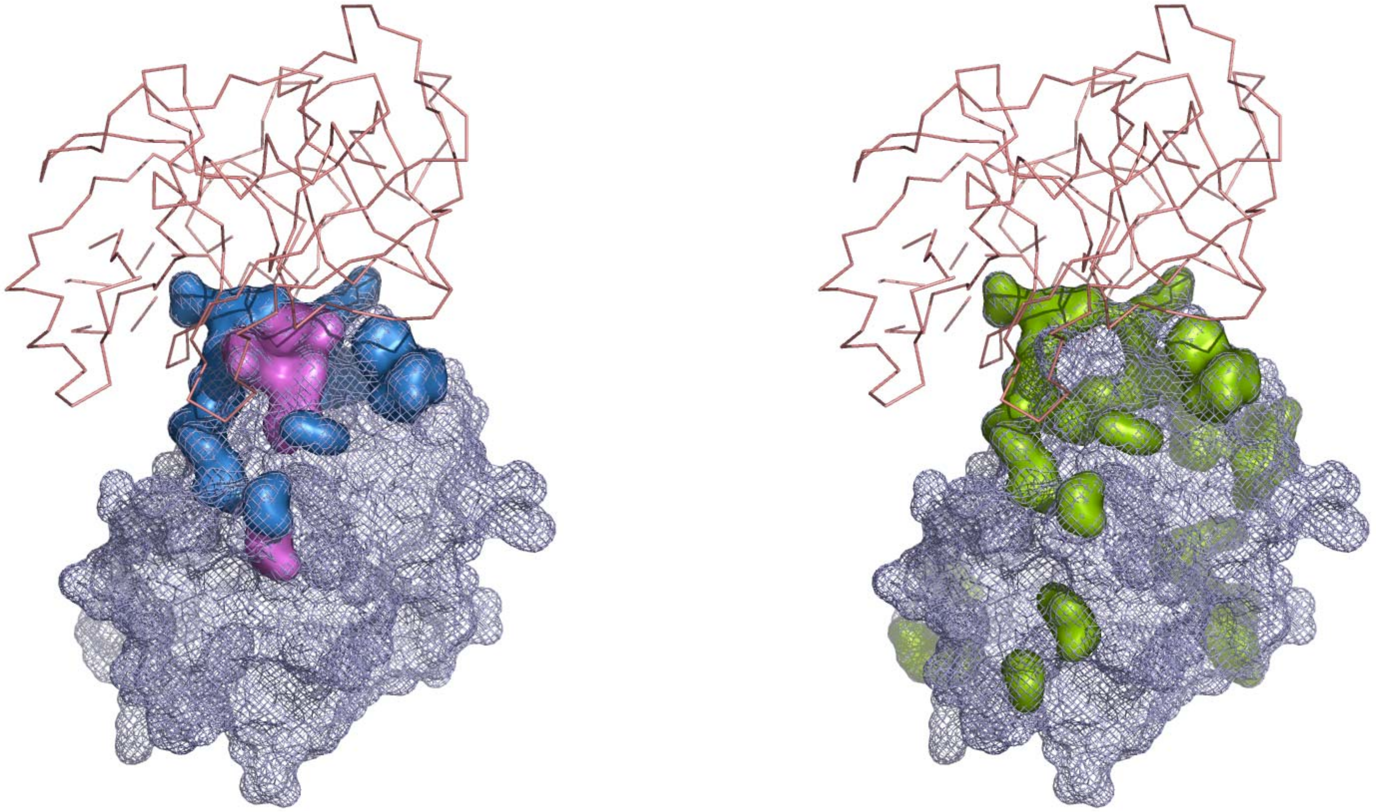


Figure 17

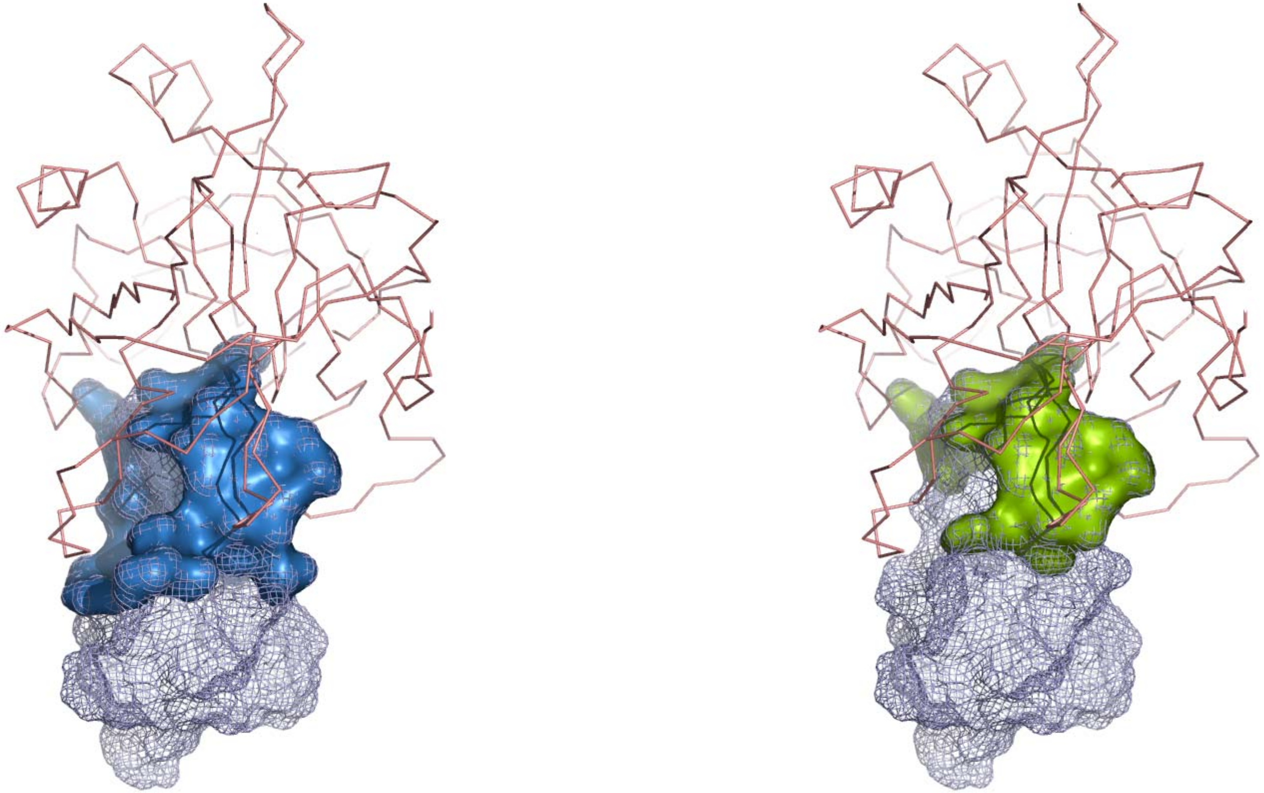


Figure 18

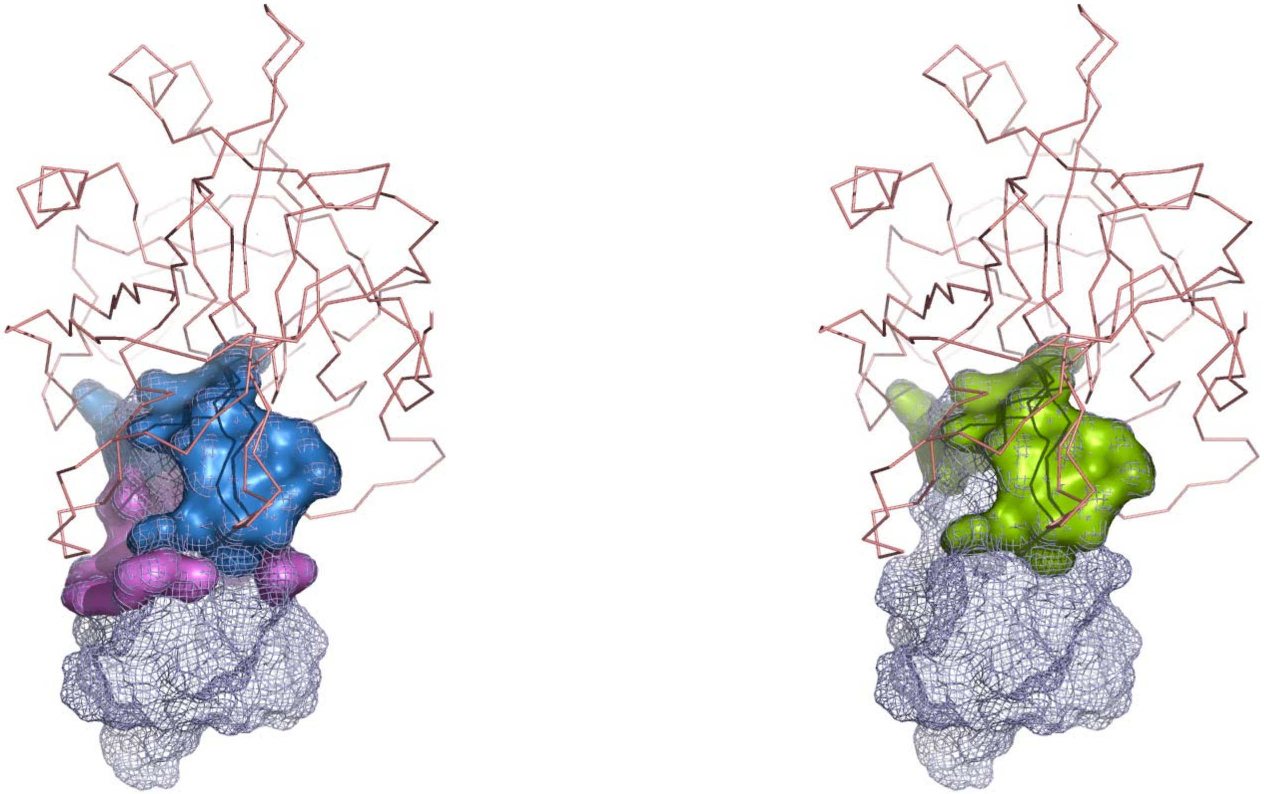
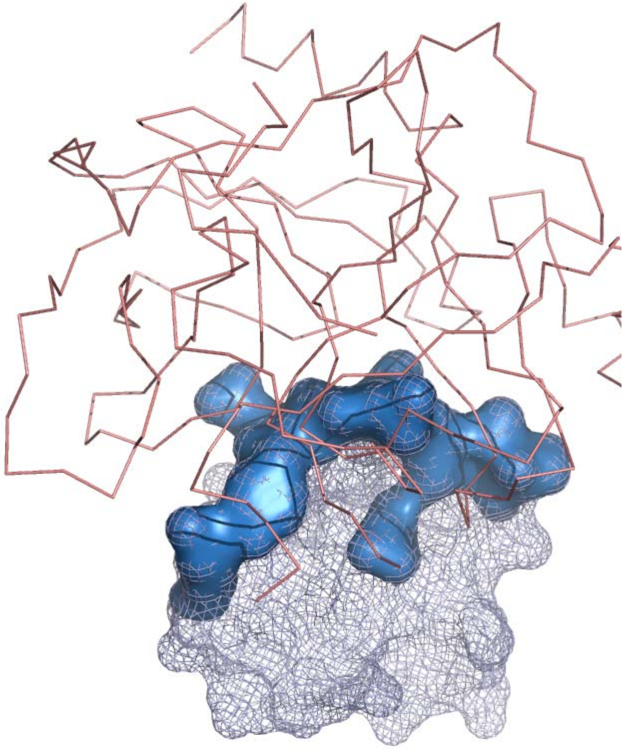


Figure 19



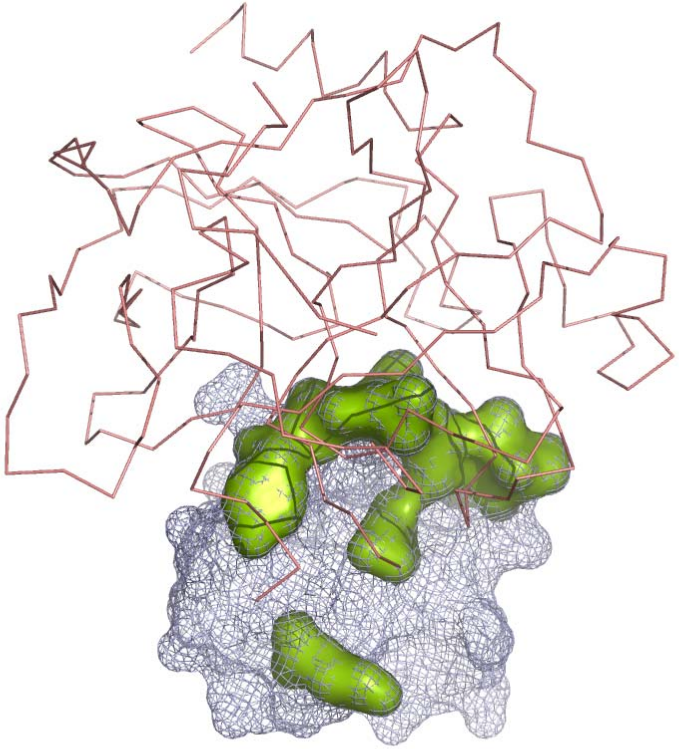


Figure 20

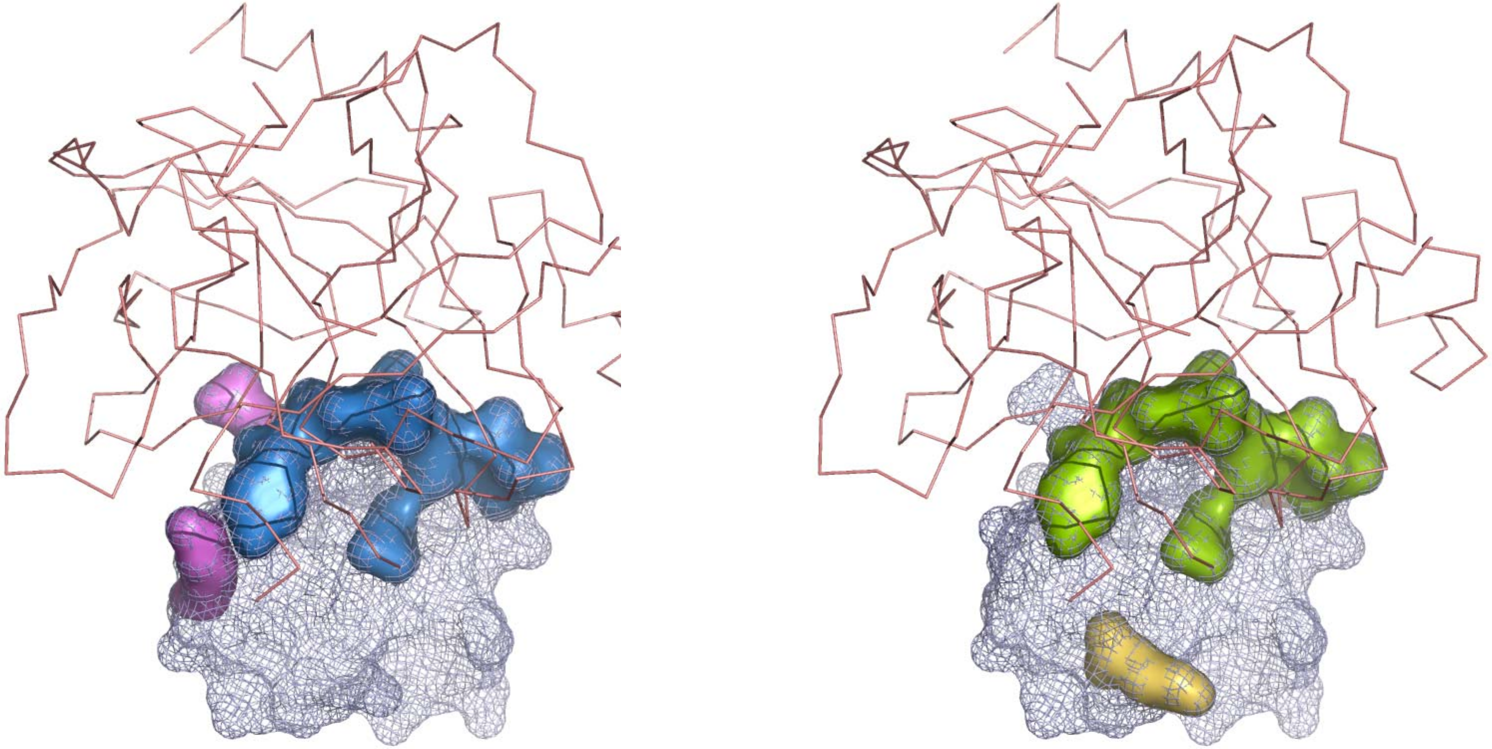


Figure 21

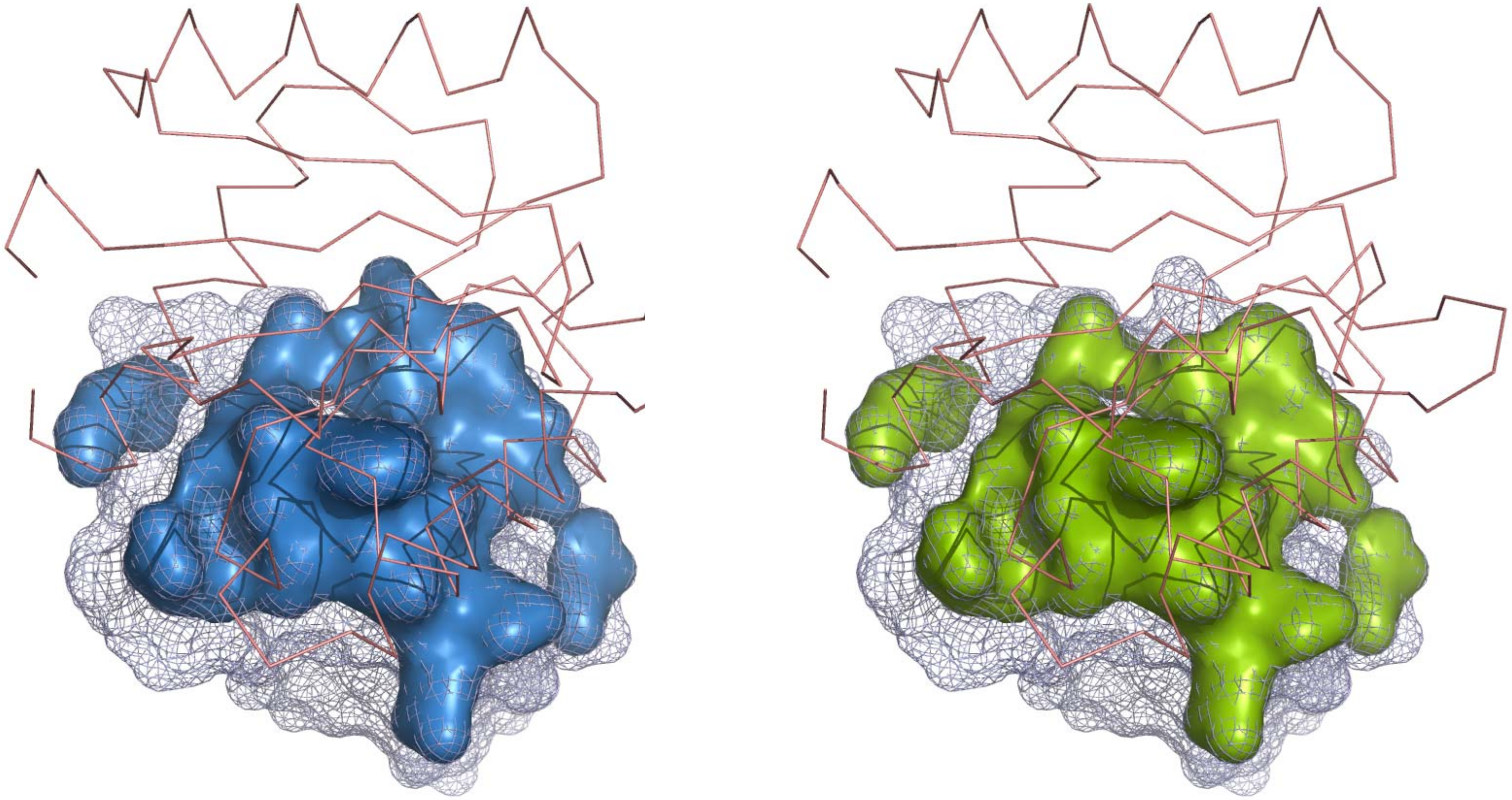


Figure 22

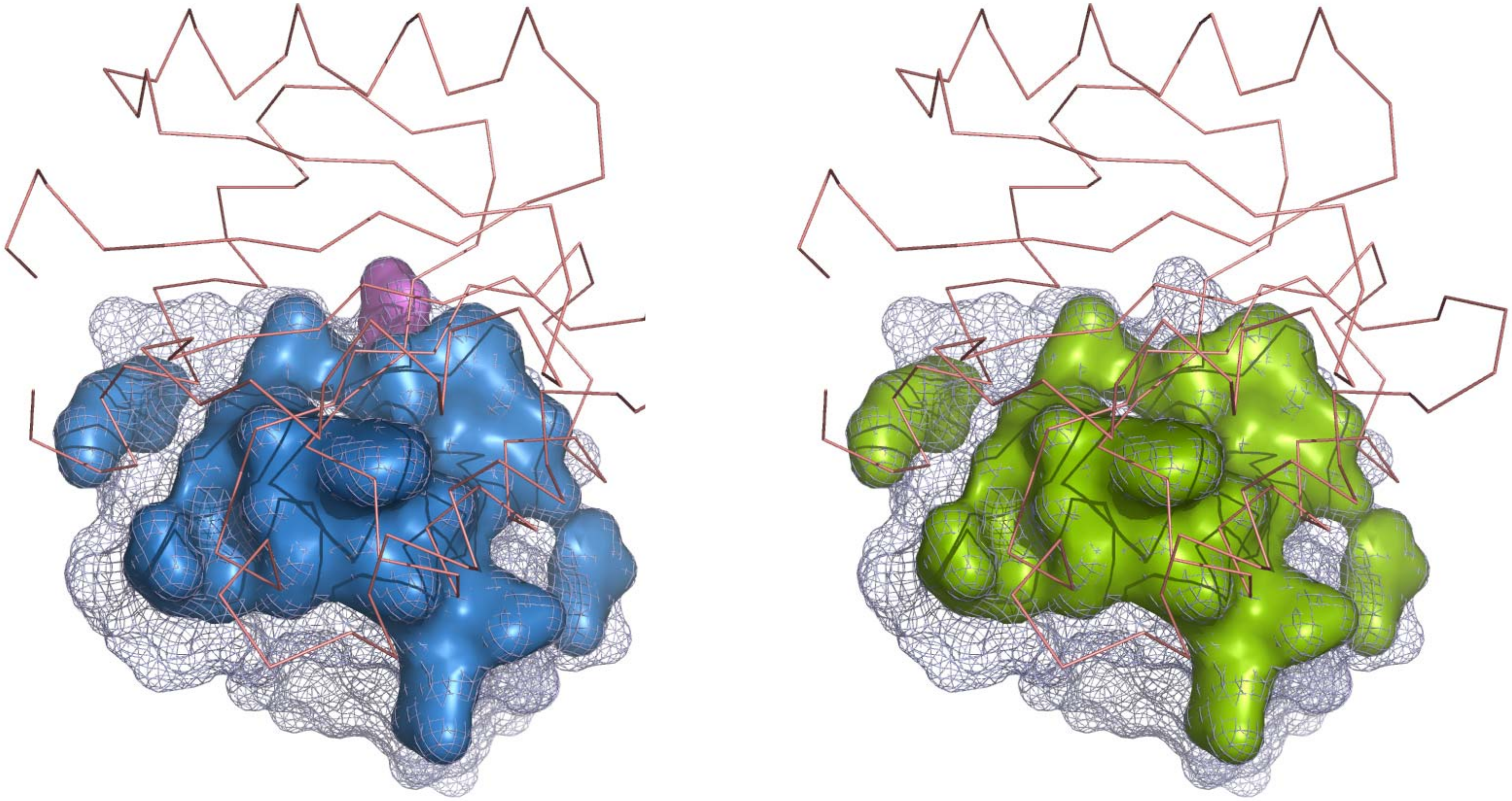


Figure 23

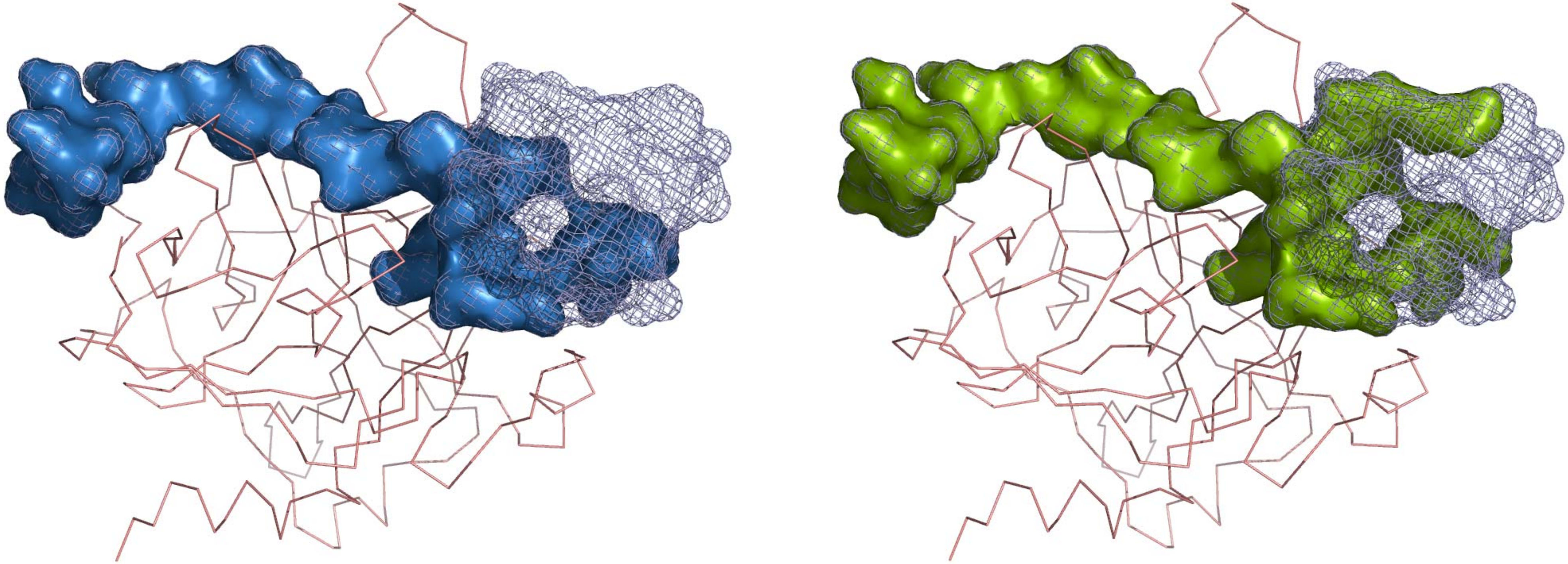


Figure 24.