Topological Autocorrelations for Prediction of Protein Conformational Stability and Kinase and Protease Inhibitions

Michael Fernandez Llamosa
(Student ID: 08791007)

Supervisor
Prof. Akinori Sarai

DEPARTMENT OF BIOSCIENCE AND BIOINFORMATICS
KYUSHU INSTITUTE OF TECHNOLOGY
IIZUKA, JAPAN

Table of Contents

Acknowledgments

This thesis arose out of years of research that has been done since I came to Japan. By that time, I have worked with a great number of people whose contribution in assorted ways to the research and the making of the thesis deserved special mention. It is a pleasure to convey my gratitude to them all in my humble acknowledgment.

In the first place I would like to record my gratitude to Professor Akinori Sarai for his supervision, advice, and guidance from the very early stage of this research as well as giving me extraordinary experiences through out the work.

Many thanks go in particular to Dr. Shandar Ahmad and Dr. Satoshi Fujii. I am much indebted to them for their valuable advice in science discussion, supervision and furthermore. I gratefully thank my friends Faisal Zaman and Dr. Rahul Kumar for science discussions and life sharing. I would also like to thank all my friends for their support and the enjoyable moments.

I thank the Kyushu Institute of Technology (KIT) for providing me the support and equipment I have needed to produce and complete my thesis as well as the Monbukagakusho fellowship, which has funded my studies. My gratitude goes to Yuka Kobayashi and Mariko Jo that for their kind help and assistance.

Finally, I thank my family for supporting me throughout all my studies and specially my wife Mayren who have walked all this long way by my side.

Michael Fernandez Llamosa

List of abbreviations

Absorption, distribution, metabolism, excretion (ADME)

Accessible surface area (ASA)

Amino Acid Sequences Autorrelation (AASA)

Aqueous solubility (logS)

Artificial neural networks (ANNs)

Bayesian-regularized artificial neural networks (BRANNs)

Bayesian-regularized genetic neural networks (BRGNNs)

Cellular automata (CA)

Change of unfolding Gibbs free energy change (ΔΔG)

Chaos game representation (CGR)

Comparative molecular field analysis (CoMFA)

Constructive Lipophilicity（ClogP）

Empirical data-based energy function (EEEF)

European Bioinformatics Institute (EBI)

Five-Fold-Out (FFO)

Flexibility (f)

Gene Ontology (GO)

Genetic Algorithm (GA)

Genetic Algorithm-optimized Support Vector Machines (GA-SVMs)

Genetic Algorithm-optimized support vector machines (GA-SVMs)

Gibbs Free energy change of hydration for native protein (G_{hN})

Heat capacity (Cp)

Hydration heat capacity change (ΔCp_h)

Lipophilicity (logP)

Matthews's correlation coefficient (Cr)

Mean squared error (MSE)

Misclassifications of FFO crossvalidation (MC_{FFO})

National Center for Biotechnology Information (NCBI)

Number of correct prediction for class s/all prediction made for s (P(s))

Number of correct prediction for class s/observed in class s (Q(s))

Number of medium range contacts (N_m)

Partial least squares (PLS)

Physical effective energy function (PEEF)

Polarity (P)

Precision scores for positive (P(-))

Precision scores for positive class (P(+))

Proteochemometrics (PCM)

Quantitative structure-activity relationships (QSAR)

Quantum Mechanical/Molecular Mechanical (QM-MM),

Radial Basic Function (RBF)

Root MSE_{FFO} ($RMSE_{FFO}$)

Sensitivity of positive class prediction (Q(+))

Sequence Structural Fragments (SSF)

Solvent accessible surface area for native protein (ASA_N)

Specificity of positive class prediction (Q(-))

Squared correlation coefficient of FFO crossvalidation (R^2_{FFO})

Statistical potential-based effective energy function (SEEF)

Structural Fragments (SF)

Support vector machines (SVMs)

Thermodynamic transfer hydrophobicity (H_t)

Three-Fold-Out (TFO)

Unfolding Gibbs Free energy change of side-chain (ΔG_C)

Abstract

The annotation of protein structure and function from sequence and the prediction of compound's activity from sketch representations are fundamental goals in bio- and chemoinformatics. In the present study, fast and accurate predictors of protein conformational stability and kinase and protease inhibitions were built from graph representations of proteins and ligands. Firstly, Amino Acid Sequence Autocorrelation (*AASA*) vectors were computed from the Cα-carbon linear graph representation of a large dataset of protein mutants (>1000) from Protherm database. Genetic algorithm-optimized support vector machines (GA-SVM) were trained with *AASA* vectors to predict the real $\Delta\Delta G$ values with squared correlation coefficient of 0.45 and classify $\Delta\Delta G$ signs with accuracy of 80%. The stable mutants in the test set were recognized with accuracies of 70%. Secondly, *AASA* vectors and ligand's autocorrelation features were computed from the linear graph representation of kinase and protease and from 2D molecular graphs collected from ProLINT database. SVMs trained with concatenated autocorrelation matrices yielded test set accuracies > 80% for kinase and protease targets. The inhibition predictors perform homogenously along the different kinase and protease families and ligands' scaffolds. The predictors from sequences and sketch representations of ligands are online available at:

http://gibk21.bse.kyutech.ac.jp/llamosa/ddG-AASA/ddG_AASA.html

http://gibk21.bse.kyutech.ac.jp/AUTOkinI/SVMpredictor.html

http://gibk21.bse.kyutech.ac.jp/AUTOprotI/SVMpredictor.html

CHAPTER 1. INTRODUCTION

1.1 Background

A growing amount of biological data is being generated by the rapid progress of biotechnologies and the human genome project [1]. Information on metabolic pathways, protein structures, nucleic acid sequences, drug metabolism and toxicity, biomolecule interactions, biological organisms, diseases, scientific literature, and further more is publicly accessible through more than 500 databases containing different subset of biological knowledge [2, 3]. Some representative examples include diverse databases as: PubMed [4], the searchable compendium of biological literature that is maintained by the National Center for Biotechnology Information (NCBI); Ensembl [5], the database of human gene predictions that is maintained by the European Bioinformatics Institute (EBI) and the Wellcome Trust; the UCSC Genome Browser a human, mouse and rat genome browser [6] that is maintained by the University of California at Santa Cruz; FlyBase [7], the *Drosophila* research community database that is maintained by the FlyBase Consortium; WormBase [8], the *Caenorhabditis elegans* model-organism database; the Gene Ontology (GO) database [9] of gene function, process and location terms; and ChEMBLdb, a database of bioactive drug-like small molecules that is maintained by European Bioinformatics Institute [10].

Among the on-line accessible databases for structural and functional information of proteins, ProTherm: Thermodynamic Database for Proteins and Mutants [11], is the first and more comprehensive thermodynamic database available online since 1998 from the BIOINFO BANK laboratory of the Department of Bioscience and Bioinformatics of the Kyushu Institute of Technology, Japan. ProTherm includes thermodynamic and structural data on proteins and mutants along with measuring methods, experimental conditions and literature information.

Our laboratory created and maintains a database of protein-ligand interactions collected from the literature and developed as web-based public accessible database. ProLINT: Protein-Ligand Interactions includes binding data on kinases and proteases [12], sequences and structural information regarding the protein targets, structural information regarding the ligands together with experimental details and literature information.

Accessing to biological database resources is indispensable to conduct nowadays research in fields like molecular biology, genetics and medicinal chemistry. The routinary use and analysis of these information resources remains challenging to experimentalist researchers although great efforts are devoted to database integration and the development of computational tools [13]. Life-sciences scientists are now building heterogeneous models by linking analysis of different domains to elucidate the keys of their interactions for a better comprehension of the biological systems [13]. In this context, web-server and stand-alone applications provided to end-users, with different backgrounds and skill sets, gain additional value as computational applications to analyze the abundant information available. Computational biology offers tools that can provide insight into the functions of proteins based on sequence, structure and evolutionary history. Computational analysis of protein structure, function and interactions using biological databases involve several techniques such as homology modeling, molecular dynamics, docking, quantum chemistry, machine learning, quantitative-structure function relationship and so on.

1.1.1 Computational prediction of protein conformational stability

Defective protein folding is an important cause of mutation-related diseases, thus predicting protein structures and stability is a fundamental goal in molecular biology and even predicting changes induced by point mutations has immediate application in computational protein design [14-16]. Conformational stability is experimentally measured as the unfolding Gibbs free energy change (*ΔG*) (Figure 1.1).

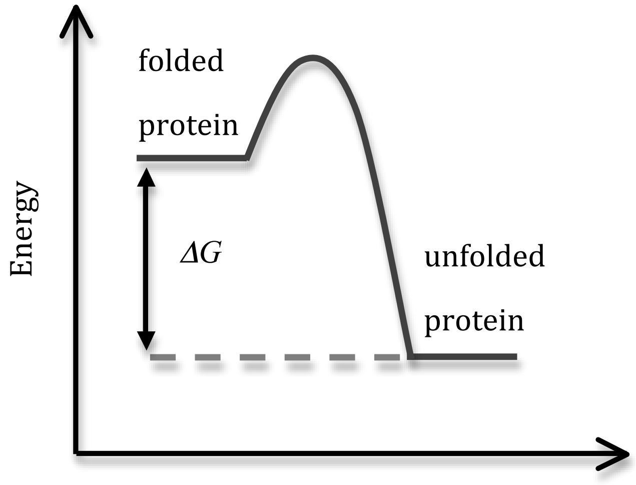


Figure 1.1. Protein unfolding Gibbs free energy change (*ΔG*).

Although free energy simulations have accurate predicted relative stabilities of point mutants [17], the computational cost of the actual methods are extremely high to test large number of mutations in protein design applications. Translation of structural data into energetic parameters is intended today by developing fast algorithms for protein energy calculations. However, the development of fast and reliable protein force-fields is a complex task due to the delicate balance between the different energy terms which contribute to protein stability. Force-fields for predicting protein stability can be divided in three main groups: physical effective energy function (PEEF) [18, 19], statistical potential-based effective energy function (SEEF) [20, 21, 22] and empirical data-based energy function (EEEF) [23, 24].

Furthermore, stability prediction studies not based on protein force-field calculations have been focused on correlations of free energy change with structural, sequence information and amino acid properties such as hydrophobicity, accessible surface area, etc. In this sense, Gromiha et al. had reported some of the seminal works in this topic [25-27]. Furthermore, empirical equations involving physical properties, have been calculated from mutant structures. Zhou and Zhou [28] reported a broad study regarding 35 proteins and 1023 mutants from which they derived a new stability scale. A "transfer free energy" scale was extracted assuming that the mutation-induced stability change is equal to the change in transfer free energy without needing any structural information.

Likewise, some X-ray structural-independent stability prediction methods have gained attention. The advantages of such methods are that they just employ amino acid sequence information for predicting protein stability and are extremely less computational intensive in comparison with free energy function methods [28, 29]. Levin and Satir [28] successfully evaluated the functional significance of mutations on hemoglobin by amino acid similarity matrixes. Frenz [29] reported a nonlinear model for predicting the stability of Staphylococcal Nuclease mutants by amino acid similarity scores. Outstanding reports of Capriotti et al. [30-32] describe predictors of the unfolding change of protein Gibbs free energy change ($\Delta\Delta G$) upon mutations by sequences and 3D structures from a dataset of more than 2000 mutants.

More recently, new predictors have been published using sequence and/or 3D structure information. Particularly, iPTREE-STAB server [33] discriminates the stability of proteins and predicts their changes upon single amino acid substitutions from amino acid sequence. Similarly, Cheng et al. have developed sequence and 3D structure-based SVM predictors [34]. In addition, Parthiban reported the prediction of protein mutant stability from distance and torsion potentials [35].

However, the efficient and accurate computational prediction of conformational stability of protein remains as a challenge. The implementation of simple and intuitive sequence-based approaches that consider the effects of multiple point mutations, insertion or deletion of residues is particularly important.

1.1.2 Kinase inhibitor therapeutic agents and affinity prediction

Death by cancer has recently increased all around the world [36]. It has been reported that the improvements in conventional cancer treatments such as surgery, radiation and cytotoxic chemotherapy will not substantially impact the clinical outcomes for cancer patients in the future [36]. Due to this, researches have focused on alternative clinical strategies such as the development of a variety of protein-targeted molecule-based cancer therapies, especially selective kinase inhibitors [36]. Kinase, alternatively known as a phosphotransferase, is a type of enzyme that transfers phosphate groups from high-energy donor molecules, such as ATP, to specific substrates. The process is referred to as phosphorylation.

Need for discovery of novel kinase inhibitors has acquired a particular significance, while has led to more basic efforts at the understanding of kinase inhibition process and methods to predict the stability of a potential kinase-inhibitor complex. There are approximately 500 kinases encoded in the human genome having potential role in cancer [37-39]. The number of available high-resolution X-ray crystal structures of kinase-inhibitor complexes has substantially increased during recent years. Structural information obtained from these complexes has become an important guide in designing selective potential kinase inhibitors. It is of utmost importance to make the best use of available structure information from these complexes in order to predict the behaviour of hypothetical complexes as an aid for inhibitor design [40-41].

Inhibition of protein kinases can be broadly classified into three categories: ATP-competitive inhibition, substrate-competitive inhibition, and allosteric inhibition. Successful treatments of chronic myeloid leukemia and gastrointestinal stromal tumor with Gleevec [42] have recently drawn much attention because of its excellent selectivity and its ability to bind to a precise inactive conformation of Abl kinase. However, the emergence of drug-resistant mutants [43] and structural studies suggest that mutations in the kinase domain cause resistance to the Abl kinase inhibitor [44]. Other studies have also shown that some inhibitors can recognize specific inactive conformation of B-Raf (1UWH) and p38 (1W83), whereas some others can inhibit the active form of Abl kinase [45]. But all of them have been shown to decrease or completely lose inhibitory activity towards some mutated kinase. In this sense, Thaimatta et al. [46], in the review of kinase inhibitors, states that the modulation of kinase activity has not been sufficiently exploited for therapeutic purposes. Inhibition of a single kinase may be insufficient to achieve a therapeutic benefit, and that promiscuous small-molecule kinase inhibitors or cocktails of inhibitors may be more promising than selective agents by targeting several kinases. In view of these facts, different computational approaches for kinase drug design need to be explored, in order to find novel, more efficient and side effects-free kinase inhibitors.

A large number of QSAR models on kinase inhibition have been reported and Tyrosine Kinase is the most studied family. In this regard, Kurup et al. [47] published a review of QSAR studies for the inhibitory activity of a chemically wide dataset towards five Tyrosine Kinases: Epidermal Growth Factor Receptor Tyrosine Kinase, Platelet-Derived Growth Factor Receptor Tyrosine Kinase, Fibroblast Growth Factor Receptor Tyrosine Kinase of Vascular Endothelial Growth Factor Receptor Tyrosine Kinase and Non receptor Tyrosine Kinase. They reported a huge amount of 40 QSAR equations using hydrophobicity, steric and electronic descriptors. The authors did not use target information but they tried to establish target-ligand interaction hypothesis by comparing quality and descriptor occurrences on the models for different inhibitor datasets on the same target or the same inhibitor dataset for different targets. The authors used very intuitive descriptors and, although the stability of target-ligands were predicted with high crossvalidation accuracies, the use and generalization of 40 models as well as their comparative interpretation is rather rough.

1.1.3 Protease inhibitor therapeutic agents and affinity prediction

Proteases are a family of enzymes representing approximately 2% of an organism proteosome that exert bioregulation, matrix remodeling, digestion, and immune response processes [48]. These enzymes account for 5–10% of the pharmaceutical targets in the current pharmaceutical market [49]. Proteases discriminate between the many possible available substrates cleave a specific protein or peptide. Subsites controlled protease specificity by assisting in the selection and orientation of a given substrate [50]. According to the architecture of enzyme active site and mechanism of hydrolysis, there are four major classes of proteases: serine proteases (which account for approximately one-third of all proteases), aspartic proteases, cysteine proteases, and metalloproteases [49, 51].

Proteases virtually occur in all biological process and their ability to catalytically turnover substrate, makes them ideal biomarkers for disease diagnostic and therapy. Infectious diseases such as malaria and Chaga's disease involve cysteine proteases, falsipain 1 [52] and cruzain [53], which help parasites in the invasion of the host cell. Similarly in HIV infection, an aspartic protease (HIV-1 protease) is responsible for the maturation of the virus [54]. In turn, proteolysis can be found to participate in five of the six processes of hallmarks of cancer [55]. A serine protease, prostate-specific antigen (PSA, human kallikrein 3), is currently used as clinical marker for prostate cancer [56]. Matrix metalloproteinases (MMPs) is a family of zinc endopeptidases involved in the connective tissue remodeling and implicated in some processes such as ovulation, embryonic growth, angiogenesis, differentiation, and healing [57]. Since any disturbance of the generally well-balanced equilibrium between the MMPs and their physiological inhibitors can provoke pathological situations such as rheumatoid and osteoarthritis, atherosclerosis, tumor development, tumor metastasis and pulmonary emphysema, MMP inhibitors have caught the interest as an important class of drugs for the development of innovative chemotherapeutics in several fields where effective treatments are lacking [58].

The development of novel diagnostic and therapeutic protease-active compounds depends on the understanding of protease's inhibition mechanism and specificity. Identification of protease substrates and hydrolytic products will contribute to elucidate the mechanisms behind the progression of disease and increase opportunities for the development of drug candidates and their interventional use [58].

The availability of three-dimensional structural information for proteases has improved this substrate-based drug design allowing receptor-based computational design. Structural information about the active site of the protease is computationally fitted into its selections of designed molecules from low molecular weight compound scaffolds. Several low molecular weight inhibitors of HIV-1 protease such as saquinavir, ritonavir, indinavir, nelfinavir, and amprenavir, currently used in humans, are among the first successful examples of receptor/structure-based designer drugs [59]. They were developed using structures of compounds bound to the active site of

HIV-1 protease and with the knowledge of inhibitors of other aspartic proteases like rennin [59].

For the last two decades, classical and 3D-QSAR approached have been extensively used to model protease inhibition [60-69]. A comprehensive review by Verma and Hansh [70] discussed a hundred of published and newly formulated QSAR models on the inhibition of various compound series against MMP-1, -2, -3, -7, -8, -9, -12, -13, and -14 were discussed in the context of the chemical–biological interactions. Similar to Kurup et al. [47], they established target-ligand interaction hypothesis by comparing the quality and descriptor occurrences on the models for different inhibitor datasets on the same target or the same inhibitor dataset for different targets. Another review on FXa inhibitors discussed 3D-QSAR studies and classical QSAR approaches on chemically diverse data sets ranging from 20 to 80 compounds [71]. The most correlating features were hydrophilicity-related properties such as ClogP with molar refractivity and sterimol parameters also important in most of the models [71]. Although the ligand activity were predicted with high accuracies, the use and generalization of several models as well as the comparative interpretation is rather rough. Additional efforts are needed to take the results of computational studies to the level of experimental accuracy, both to provide a screened set of compounds as well as predict the outcome of experiments.

1.2 Objectives

The main goal of this thesis was to develop three fast and simple computational tools, which accurately could predict the conformational stability of protein mutants and classify the inhibition affinity of ligands towards kinases and proteases. The first predictor should help on the design of new proteins and to identify disease-causing mutations. The second and the third one should allow the efficient filtering and screening of compound databases for putative potent inhibitors. To accomplish this objective the following specific tasks were developed:

1.2.1- Prediction of protein conformational stability

- To build a protein conformational stability dataset by collecting protein sequences, experimental values of change of free energy change (*ΔΔG*) upon mutation and structural and experimental information from Protherm database [11].
- To calculate autocorrelation features from the Cα-carbon linear graph representation of mutants sequence.
- To optimize support vector machines to predict protein conformational stability from autocorrelation features using genetic algorithm-based hyperparameter optimization and feature selection.
- To analyze general predictor performance and evaluate the accuracy for different types of mutations.
- To analyze the optimum correlating properties in terms of the contribution to the conformational stability.
- To implement a web-server application to predict the conformational stability of protein sequence through the world-wide-web.

1.2.2- Prediction of kinase and protease inhibitions

- To build two inhibition datasets for kinase and protease by collecting protein sequences, ligands structures and experimental measures of inhibition affinity from ProLINT database [12].
- To calculate topological autocorrelation features from the weighted linear graph representation of targets and 2D graphs of the inhibitors.
- To analyze target similarity based on the distribution of the high

affinity inhibitors.

- To optimize support vector machines to predict inhibition using concatenated matrices of autocorrelation features.
- To analyze the predictor performance and accuracy for different target families and inhibitor scaffolds.
- To implement two web-server applications to predict inhibitor affinity towards different kinase and protease sequences.

1.3 Thesis Outline

The thesis is organized in six chapters. The Chapter 2 discusses datasets and computational methods used in the present study. The Chapter 3 describes the computation of Amino Acids Sequence Autocorrelation (*AASA*) vectors from the Cα-carbon linear graph representation of sequences and the implementation of support vector machines (SVM) and genetic algorithm (GA) hybrid approach to predict real values and signs of unfolding $\Delta\Delta G$ of protein upon mutations. In Chapter 4, *AASA* vectors were calculated on kinase sequences and combined with autocorrelation features from ligand 2D graphs to train SVM models of kinases inhibition. The Chapter 5 deals with the same framework developed in Chapter 4 but now applied to predict affinity of ligands towards proteases. Chapter 6 summarizes the results and discusses the prospective based on the present study.

1.4 References

1. Aldhous PM: **Human genome project. Database goes on–line**. *Nature* 1990, **347**:9.

2. Baxevanis AD: **The molecular biology database collection. 2003 update.** *Nucleic Acids Res* 2003, 31:1–12.

3. Discala C, et al.: **DBcat: a catalog of 500 biological databases**. *Nucleic Acids Res* 2000, **28**:8–9.

4. Wheeler DL, et al.: **Database resources of the National Center for Biotechnology Information: 2002 update**. *Nucleic Acids Res*. 2002, **30**:13–16.

5. Hubbard T, et al. **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**:38–41.

6. Kent WJ, et al.: **The human genome browser at UCSC**. *Genome Res* 2002, **12**:996–1006.

7. The FlyBase Consortium: **The FlyBase database of the Drosophila genome projects and community literature**. *Nucleic Acids Res* 2002, **30**:106–108.

8. Harris TW, et al.: **WormBase: a cross-species database for comparative genomics**. *Nucleic Acids Res* 2003, **31**:133–137.

9. The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology**. *Nat Genet* 2000, **25**:25–29.

10. Strauss S: **Pharma embraces open source models**. *Nat Biotechnol* 2010, **28**:631–4.

11. Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34**:D204–D206.

12. Ahmad S, Kitajima K, Selvaraj S, Kubodera H, Sunada S, An J-H, Sarai A: **Protein-Ligand Interactions: ProLINT Database and QSAR Analysis.** *Genome Inform* 2003, **14**:537–538.

13. Curcin V, Ghanem M , Guo Y: **Web services in the life sciences**. *Drug Discov Today* 2005, **10**:865–871.

14. a) Saven J: **Combinatorial Protein Design**. *Curr Opin Struct Biol* 2002, **12**:453–458. b) Mendes J, Guerois R, Serrano L: **Energy Estimation in Protein Design**. *Curr Opin Struct Biol* 2002, 12:441–458.

15. Bolon DN, Marcus JS, Ross SA, Mayo, SL: **Prudent Modeling of Core Polar Residues in Computational Protein Design**. *J Mol Biol* 2003, **329**: 611–622.

16. Looger LL, Dwyer MA, Smith JJ, Helling HW: **Computational Design of Receptor and Sensor Proteins with Novel Functions**. Nature 2003, **423**: 185–190.

17. Dang LX, Merz KM, Kollman, PA: **Free-energy Calculations on Protein Stability: Thr-1573Val-157 Mutation of T4 Lysozyme**. *J Am Chem Soc* 1989, **111**:8505–8508.

18. Lee C, Levitt, M: **Accurate Prediction of the Stability and Activity Effects of Site- Directed Mutagenesis on a Protein Core**. *Nature* 1991, **352**:448–451.

19. Lee C: **Testing Homology Modeling on Mutant Proteins: Predicting Structural and Thermodynamic Effects in the Ala98-Val Mutants of T4 Lysozyme**. *Fold Des* 1995, **1**:1–12.

20. Lazaridis T, Karplus M: **Effective energy functions for protein structure prediction**. *Curr Opin Struct Biol* 2000, **10**:139–145.

21. Topham CM, Srinivasan N, Blundell TL: **Prediction of the Stability of Protein Mutants Based on Structural Environment-dependent Amino Acid Substitution and Propensity Tables**. *Protein Eng* 1997, **10**:7–21.

22. Gilis D, Rooman M: **Prediction of Stability Changes upon Single site Mutations Using Database-Derived Potentials**. *Theor Chem Acc* 1999, **101**: 46–50.

23. Lacroix E, Viguera AR, Serrano L: **Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters**. *J Mol Biol* 1998, **284**:173 –191. b) Munoz V, Serrano L: **Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm–Bragg and Lifson–Roig formalisms**. *Biopolymers* 1997, **41**:495–509.

24. Guerois R, Nielsen JE, Serrano L: **Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations**. *J Mol Biol* 2002, **320**:369 –387.

25. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A: **Relationship Between Amino Acid Properties and Protein Stability: Buried Mutations.** *J Prot Chem* 1999**, 18**: 565–578.

26. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A: **Role of Structural and Sequence Information in the Prediction of Protein Stability Changes: Comparison between Buried and Partially Buried Mutations**. *Protein Eng* 1999, **12**:549 –555.

27. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A: **Importance of Surrounding Residues for Protein Stability of Partially Buried Mutations**. *J Biomol Struct Dyn* 2000, **18**:1–16.

28. Levin S, Satir BH: **POLINA: Detection and Evaluation of Single Amino Acid Substitutions in Protein Superfamilies**. *Bioinformatics* 1998, **14:**374–375.

29. Frenz CM: **Neural Network-Based Prediction of Mutation-Induced Protein Stability Changes in Staphylococcal Nuclease at 20 Residue Positions**. *Proteins* 2005, **59**:147–151.

30. Capriotti E, Fariselli P, Casadio R: **A neural-network-based method for predicting protein stability changes upon single mutations**. *Bioinformatics* 2004, **20**:63–68.

31. Capriotti E, Fariselli P, Calabrese R, Casadio R: **Prediction of protein stability changes from sequences using support vector machines**. *Bioinformatics* 2005, **21**:54 –58.

32. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure**. *Nucleic Acids Res* 2005, **33**:306–310.

33. Huang L-T, Gromiha MM, Ho S-Y: **iPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations**. *Bioinformatics* 2007, **23**:1292–1293.

34. Cheng J, Randall A, Baldi P: **Prediction of protein stability changes for single-site mutations using support vector machines**. *Proteins* 2006, **62**: 1125–1132.

35. Parthiban V, Gromiha MM, Hoppe C, Schomburg D: **Structural analysis and prediction of protein mutant stability using distance and torsion potentials: Role of secondary structure and solvent accessibility**. *Proteins* 2006, **66**:41–52.

36. Settleman, J: **Mutated kinases as targets for cancer drugs**. *Drug Discov Today: Disease Mechanisms* 2005, **2**:139–144.

37. Manning DB, Whyte R, Martinez T, Hunter T, Sudarsanam S: **The protein kinase complement of the humangenome**. *Science* 2002, **298**:1912–1934

38. Cohen P: **Protein kinases–the major drug targets of the twenty-first century**. *Nat Rev Drug Discov* 2002, **1**:309–315.

39. Faivre S, Djelloul S, Raymond E: **New paradigms in anticancer therapy: targeting multiple signaling pathways with kinase inhibitors**. *Semin Oncol* 2006, **33**:407–420

40. Vieth M, Higgs RE, Robertso DH, Shapiro M, Gragg EA, Hemmerle H: **Kinomics-structural biology and chemogenomics of kinase inhibitors and targets**. *Biochim Biophys Acta* 2004, **1697**:243–257

41. Fedorov M, Sundström B, Marsden B, Knapp S: **Insights for the development of specific kinase inhibitors by targeted structural genomics**. *Drug Discov Today* 2007, **12**:362–365.

42. Bogoyevitch MA, Fairlie, DP: **A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding**. *Drug Discov Today* 2007, **12**:622–623.

43. Hochhaus A, La Rosee P: **Imatinib therapy in chronic myleogenous leukemia: Strategies to avoid and overcome resistance**. *Leukemia* 2004, **18**:1321–1331.

44. a) Nagar B, Bornmann WG, Pellicena P, Schindler T, Veach DR, Miller WT, Clarkson B, Kuriyanet J: **Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571)**. *Cancer Res* 2002, **62**:4236–4243. b) Schindler T, Bornmann W, Pellicena P, Miller TW, Clarkson B, Kuriyan J: **Structural mechanism of STI-571 inhibition of Abelson tyrosine kinase**. *Science* 2000, **289**:1938–1942.

45. Manley WP, Cowan-Jacob WS, Mestan J: **Advances in the structural biology, design and clinical development of Bcr-Abl kinase inhibitors for the treatment of chronic myeloid leukaemia**. *Biochim Biophys Acta* 2005, **1754**:3–13. b) Golas MJ, Arndt K, Etienne C, Lucas J, Nardin D, Boschelli DH, Boschelli F: **SKI-606, a 4-anilino-3-quinolinecarbonitrile dual inhibitor of Src and Abl kinases, is a potent antiproliferative agent against chronic myelogenous leukemia cells in culture and causes regression of K562 xenografts in nude mice**. *Cancer Res* 2003, **63**:375–381. c) Lombardo LJ, Lee FY, Chen P, Norris D, Barrish JC, Behnia K, Castaneda S, Cornelius LA, Das J, Doweyko AM, Fairchild C, Hunt JT, Inigo I, Johnston K, Kamath A, Kan D, Klei H, Marathe P, Pang S, Peterson R, Pitt S, Schieven GL, Schmidt RJ, Tokarski J, Wen ML, Wityak J, Borzilleri RM: **Discovery of N-(2-Chloro-6-methylphenyl)-2-(6-(4-(2-hydroxy- ethyl)-piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (BMS-354825), a dual Src/Abl kinase inhibitor with potent antitumor activity in preclinical assays**. *J Med Chem* 2004, **47**:6658–6661.

46. Thaimattam R, Banerjee R, Miglani R, Iqbal J: **Protein Kinase Inhibitors: Structural Insights Into Selectivity**. *Curr Pharm Des* 2007, **13**:2751–2765.

47. Kurup A, Garg R, Hansch C: **Comparative QSAR Study of Tyrosine Kinase Inhibitors**. *Chem Rev* 2001, **101**:2573–2600.

48. Rawlings ND, Morton FR, Barrett AJ: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2006, **34**:D270–D272.

49. Salisbury CM, Ellman JA: **Rapid Identification of Potent Nonpeptidic Serine Protease Inhibitors**. *ChemBioChem* 2006, **7**:1034–1037.

50. Perona JJ, Craik CS: **Structural basis of substrate specificity in the serine proteases**. *Protein Sci* 1995, **3**:337–360.

51. Schechter I, Berger A: **On the size of the active site in proteases. I. Papain**. *Biochem Biophys Res Commun* 1967, **27**:157–162.

52. Greenbaum D, Baruch A, Grainger M, Bozdech Z, Medzihradszky K, Engel J, Holder T, DeRisi J, Bogyo M: **A role for the cysteine protease falcipain 1 in host cell invasion by the malaria parasite, Plasmodium falciparum.** *Science* 2002, **298**:2002–2006.

53. Kumar A, Kumar K, Korde R, Puri SK, Malhotra P, Chauhan VS: **Falcipain-1, a Plasmodium falciparum Cysteine Protease with Vaccine Potential**. *Infect Immun* 2007, **75**:2026–2034.

54. Navia MA, Fitzgerald PMD, McKeever BM, Leu CT, Heimbach JC, Herber WK, Sigal IS, Darke PL, Springer JP: **Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1**. *Nature* 1989, **337**:615–620.

55. Hanahan D, Weinberg R. **The hallmarks of cancer**. *Cell* 2000, **100**:57–70.

56. Stenman UH, Leinonen J, Alfthan H, Rannikko S, Tuhkanen K, Alfthan O. **A Complex between Prostate-specific Antigen and α1-Antichymotrypsin Is the Major Form of Prostate-specific Antigen in Serum of Patients with Prostatic Cancer: Assay of the Complex Improves Clinical Sensitivity for Cancer**. *Cancer Res* 1991, **51**:222–226.

57. Baker AH, Edwards DR, Murphy G: **Metalloproteinase inhibitors: biological actions and therapeutic opportunities**. *J Cell Sci* 2002, **115**:3719–3727.

58. Diamond SL: **Methods for mapping protease specificity**. *Curr Opin Chem Biol* 2007, **11**:46–51

59. Patick AK, Potts KE: **Protease Inhibitors as Antiviral Agents.** *Clin Microbiol Rev* 1998, **11**:614–627.

60. Gupta SP, Babu MS, Sowmya S: **A Quantitative Structure-Activity Relationship Study on Some Sulfolanes and Arylthiomethanes Acting as HIV-1 Protease Inhibitors**. *Bioorg Med Chem* 1998, **6**:2185–2192.

61. Huang X, Xu L, Luo X, Fan K, Ji R, Pei G, Chen K, Jiang H: **Elucidating the Inhibiting Mode of AHPBA Derivatives against HIV-1 Protease and Building Predictive 3D-QSAR Models**. *J Med Chem* 2002, **45**:333–343.

62. Katritzky AR, Oliferenko A, Lomaka A, Karelson M: **Six-Membered Cyclic Ureas as HIV-1 Protease Inhibitors: A QSAR Study Based on CODESSA PRO Approach**. *Bioorg Med Chem Lett* 2002, **12**:3453–3457.

63. Patankar SJ, Jurs PC: **Classification of HIV protease inhibitors on the basis of their antiviral potency using radial basis function neural networks**. *J Comput Aided Mol Des* 2003, **17**:155–71.

64. Fernández M, Caballero J: **Modeling of Activity of Cyclic Urea HIV-1 Protease Inhibitors using Regularized-Artificial Neural Networks**. *Bioorg Med Chem* 2006, **14**:280–294.

65. Yang XG, Lv W, Chen YZ, Xue Y: **In silico prediction and screening of gamma-secretase inhibitors by molecular descriptors and machine learning methods**. *J Comput Chem* 2010, **30**:1249–58.

66. Fernández M, Caballero J, Tundidor-Camba A: **Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino] acetamide derivatives as matrix metalloproteinase inhibitors**. *Bioorg Med Chem* 2006, **14**:4137–4150.

67. Fernández M, Caballero J: **QSAR Modeling of matrix metalloproteinase inhibition by N-Hydroxy-α-henylsulfonylacetamide derivatives**. *Bioorg Med Chem* 2007, **15**:6298–6310.

68. Fernandez M, Fernandez L, Caballero J, Abreu JI, Reyes G: **Proteochemometric Modeling of the Inhibition Complexes of Matrix Metalloproteinases with N-Hydroxy-2-[(Phenylsulfonyl)Amino]Acetamide Derivatives Using Topological Autocorrelation Interaction Matrix and Model Ensemble Averaging**. *Chem Biol Drug Des* 2008, **72**: 65–78

69. Tuccinardi T, Ortore G, Santos MA, Marques SM, Nuti E, Rossello A, Martinelli A: **Multitemplate alignment method for the development of a reliable 3D-QSAR model for the analysis of MMP3 inhibitors**. *J Chem Inf Model* 2009, **49**:1715–1724.

70. Verma RP, Hansch C: **Matrix metalloproteinases (MMPs): Chemical–biological functions and (Q)SARs**. *Bioorg Med Chem* 2007, **15**:2223–2268.

71. Kontogiorgis CA, Hadjipavlou-Litina, D: **Current Trends in Quantitative Structure Activity Relationships on FXa inhibitors: Evaluation and Comparative Analysis**. *Med Res Rev* 2004, **24**:687–747.

CHAPTER 2. DATASETS AND COMPUTATIONAL METHODS

2.1. Protein Mutants Dataset

A non-redundant version of the mutant list previously reported by Capriotti et al. [1] was collected from the Protherm database [2] according to the following constrains:

1) $\Delta\Delta G$ values have been experimentally determined and reported in the database.
2) the data is related to single point mutations (non multiple mutations were taken into account).

After filtering and removing redundant entries, a total of 1383 non-redundant single point mutants were used as training set.

A test set was prepared including non-redundant double and multiple-point mutations in Protherm [2] and single point mutations added to the database from Cappriotti's selection date (February, 2004) [2] until September, 2007. The test dataset was collected according to the following constraint:

1) $\Delta\Delta G$ values have been experimentally determined and reported in the database.
2) single mutants in Capriotti's dataset were not considered.

After filtering we gathered a test set including non-redundant 222 single, 277 double and 144 multiple-point mutations corresponding to 22, 43 and 18 proteins, respectively.

2.2. Kinase Inhibition Dataset

The kinase inhibition data: a total of 8235 inhibitors for 95 sequences of kinase, was obtained from our in-house manually-curated and annotated protein-ligand interaction data in ProLINT database [3]. Annotations include comprehensive information about experimentally determined thermodynamic, structural, clinical and activity parameters. Kinase sequences were retrieved from UniProt database [4] and added to the kinase inhibition dataset. Instant JChem software [5] was used for chemical database management. In many cases, ProLINT data do not contain or have ambiguous values for some of the parameters for a given interaction, as it may have not been reported in the corresponding literature source. The dataset was therefore filtered according to the following criteria:

1. Inhibitors reporting IC_{50} (81% of kinase entries in ProLINT contain IC_{50}, 5.7% for Ki and 13% for percent of inhibitor activity).

2. Inhibitors fulfilling the mass bioavailability constraint (molecular weight < 500 g/mol).

3. Inhibitors reporting unambiguous sequence information for the kinase on which experiment was performed.

After the filtering process, the redundant entries were removed and finally a dataset of 3595 nonredundant inhibition complexes (different ligand-target pairs) of 2233 unique inhibitors with 62 kinases from 19 kinase families was selected (data available from the authors upon request). Inhibition complexes were labelled into two classes according to the affinity threshold of 1 μM; "stable" class ($IC_{50} < 1$ μM) and "unstable" class ($IC_{50} > 1$ μM) yielding 1200 stable and 2395 unstable complexes.

2.3. *Protease Inhibition Dataset*

The protease inhibition data: a total of 5048 inhibitors for 52 sequences of proteases, were obtained from our in-house manually-curated and annotated protein-ligand interaction data in ProLINT database [3]. Annotations include comprehensive information about experimentally determined thermodynamic, structural, clinical and activity parameters. Protease sequences were retrieved from UniProt database [4] and added to the protease inhibition dataset. Instant JChem software [5] was used for chemical database management. The dataset was filtered according to the following criteria:

1. Inhibitors reporting *Ki* (40% of proteases entries in ProLINT contain *Ki*, 33% for IC_{50} and 27% for other measures of inhibitor activity).
2. Inhibitors reporting unambiguous sequence information for the protease on which experiment was performed.

After the filtering process, the redundant entries were removed and finally a dataset of 1706 nonredundant inhibition complexes (different ligand-target pairs) of 739 unique inhibitors with 32 proteases from 9 protease families were selected. Inhibition complexes were labelled into two classes according to the affinity threshold of 0.1 μM; "high affinity" class ($Ki < 0.1$ μM) and "low affinity" class ($Ki > 0.1$ μM) yielding 718 low affinity and 988 high affinity complexes.

2.4. *Representations of Protein Sequences as Linear Graphs*

Machine learning algorithms have been applied to several problems in computational biology. However, the successful application of pattern recognition techniques is tightened by the availability of protein feature vectors properly encoding structural information from sequences and/or crystal structures. Besides this fact, the scarce information provided by the string sequences means there is a necessity to develop alternative representations of the primary structure of nucleic acids and proteins and to implement novel frameworks for similarity studies.

More than one decade ago, several authors started representing biological sequences in continuous coordinate spaces [6-12]. The basic idea is to define trajectories in the continuous space conserving the statistical properties of the sequences of nucleotides in nucleic acids, aminoacids in proteins and repeated nucleotide sequences in Multi Locus Sequence Typing [13]. Ideally, the coordinate position of each unit in a sequence would uniquely encode for both its identity and its context, i.e. the identity of its neighbors [14] and should be scale-independent.

Based on CGR, different representations of DNA and protein sequences have been described**.** CGR of DNA sequence was reported by Jeffrey [9] and further systematized to any discrete sequence of regular elements [15, 16]. This technique was applied to analyse primary structures of protein sequences using a CGR template of 20 attractors (amino acids) [11]. 2D graphical representations of proteins starts at the center of the circle following the amino acid sequence by moving half way towards the corresponding amino acid, similar to the scheme of Jeffrey's CGR of DNA [9]. Randic et al. [14] described a highly compact graphical and numerical characterization of proteins constructed inside a unit "magic circle" with the 20 amino acids positioned at equal distances along its circumference. The sequences of proteins in a database were plotted inside a regular 20-vertex polygon, irrespectively of their functions or origins, looking for occurrence frequencies of special sequence motifs in the whole protein database. Yu et al. [17] reported 2D map representation of protein sequences based on detailed hydrophobic-polar HP model. The 20 different kinds of amino acids were divided into four classes: non-polar, negative polar, uncharged polar and positive polar. This method was used to develop a highly accurate linear discriminant function for the recognition of polygalacturonases from a dataset of protein sequences [18], type III Rnases [19] and protein targets of *Leishmania* parasites [20].

A 3D pseudo-folding representation proposed by Bai and Wang and derived from CGR [21] used a 3D Cartesian coordinate system. In this representation, proteins are shown in the interior of regular dodecahedron centered at origin of the Cartesian coordinate system and one vertex at the point (0, 0, 1) that is circumscribed inside a unit "magic sphere". At the dodecahedron vertexes were positioned 20 amino acids. We applied the "moving across the sequence" scheme, reported by Jeffrey [9], to calculate 3D coordinates for amino acid residues in a sequence [22]. This representation successfully recognized stable proteins [22] and successfully classified calcium channel proteins into electro-physiological classes [23].

Recently, protein sequences were represented as networks where the amino acids are the vertexes (nodes), connected in a specific sequence by the peptide bonds. This star graph is a special case of trees with N vertexes, where one has N-1 degrees of freedom and the remaining N-1 vertexes have just a single degree of freedom [24, 25]. Visualization of biological sequences has been also developed using cellular automata (CA) [26]. Sequences were translated into a set of dynamical systems in which space and time are discrete. Transforming the symbolic sequence into digital codes through optimal space-time evolution rules generated a unique CA image. Xiao et al. [27] showed that this representation depicts some important features originally hidden in a long and complicated biological sequence, such as the characteristic vector of a sequence.

A very simple and straightforward representation consists of depicting sequence as linear graph in which nodes represent Cα atoms of amino acid residues (Figure 2.1) and weight values (properties) are assigned to each node (residue). In fact, this approach can be considered as an extension of the popular and widely used 2D graph representation of low molecular weight compounds to polymeric macromolecules taking amino acids residues as structural units. Different sets of amino acid/residues properties from AAindex database [28] have been used for generating weighting patterns for such graphs. Similarly, the well-known z-scores from principal component analysis of amino acid/residues properties have been used as weights [29]. A variety of feature vectors have been calculated using these representations, based on graph index invariants and topological indexes. Amino Acid Sequences Autorrelation (*AASA*) [30, 31], protein linear indexes of the 'macromolecular pseudo graph Cα-atom adjacency matrix [32], pseudo-amino acid composition [33, 34] as well as full sets of topological and markovian-derived descriptors [35, 36] have been proposed for encoding properties distributions and quasi-order information in protein sequence. This representation has been successfully applied, either explicitly or implicitly, to predict subcellular locations [33], secondary structure [34] of proteins, the conformational stabilities of single protein mutants [30, 1] and receptor affinities [37].

In addition to its simplicity and success, this representation is the most suitable to combine macromolecular targets and low molecular weight ligands into a simple and sound representation. Regarding this, sequences extracted from ProTherm [2] and ProLINT [3] databases were represented as weighted linear graphs in our studies. As weights for sequence residues were employed, physicochemical and conformational amino acid/residues properties (Table 2.1 and 2.2 Appendix) selected from the AAindex database [28].

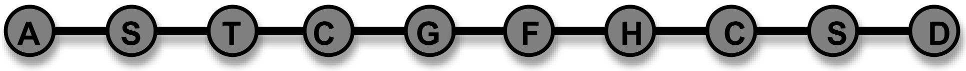


Figure 2.1. Schematic representation as linear graph of a hypothetic decapeptide with sequence ASTCGFHCSD.

2.5. Calculation of Structural Features of Proteins and Inhibitors

The *in-silico* study of biomolecule structures and interactions constitutes a parallel and complementary approach to experimental research, which have used molecular dynamics simulations, Quantum Mechanical/Molecular Mechanical (QM-MM), docking, and Quantitative-Structure Activity Relationships (QSAR) computational techniques [38]. In QSAR studies (Figure 2.2), structural features are correlated to activity/properties real values or classification schemes using mathematical functions that range from simple linear regressions to sophisticate machine learning implementations. This approach includes a large number of models for the prediction of inhibitor affinity values [39, 40].

QSAR and docking techniques have yielded the most of the models for predicting enzyme-inhibitor binding affinities [41]. Even if protein targets are closely related or correspond to the same protein family, different targets should constitute individual training sets to model multiple target-ligand systems. In this sense, this technique produces a huge amount of models, each one applicable to each target, to model affinities towards multiple targets. This fact makes difficult the model's interpretation and generalization, even for closely related targets. Additionally, docking studies employ 3D structures from targets and ligands for generating interaction scores for ligand-target binding conformations. The main uses of docking are to select hits in virtual library screening and to evaluate 3D binding modes. This technique is very convenient to preselect "true binders" for chemical library from virtual library. A detailed description of target and ligand 3D structures is needed for docking. However, only for closely related target some generalizations can be made if very good alignments are available and ligands are also very familiar [41]. Generalized models for enzyme inhibition that include ligands from different scaffolds will be difficult to build by docking.

Proteochemometrics (PCM) a variant of QSAR analysis proposed by Wikberg [42], originates from chemometrics, the mathematical methods to analyze chemical data. Based on a specific structure representation, PCM models can describe the properties of macromolecules (such as proteins) as well as the interactions between them and a series of ligands. These models are useful for predicting the properties of new proteins as well as the affinities of new proteins for their ligands. Similarly, one PCM model can predict the affinity of new ligands towards a group of related targets. A PCM experiment is typically described by three descriptor blocks; the ligand descriptor, protein descriptor, and ligand-protein cross-term blocks. A vector of variables of ligand descriptors characterizes each ligand. Similarly, each protein is described by protein descriptors. Depending on the problem, one or more descriptor blocks can be discarded. In our study, the cross-term blocks were discarded since nonlinearity was automatically incorporated into the models by nonlinear approximation methods.

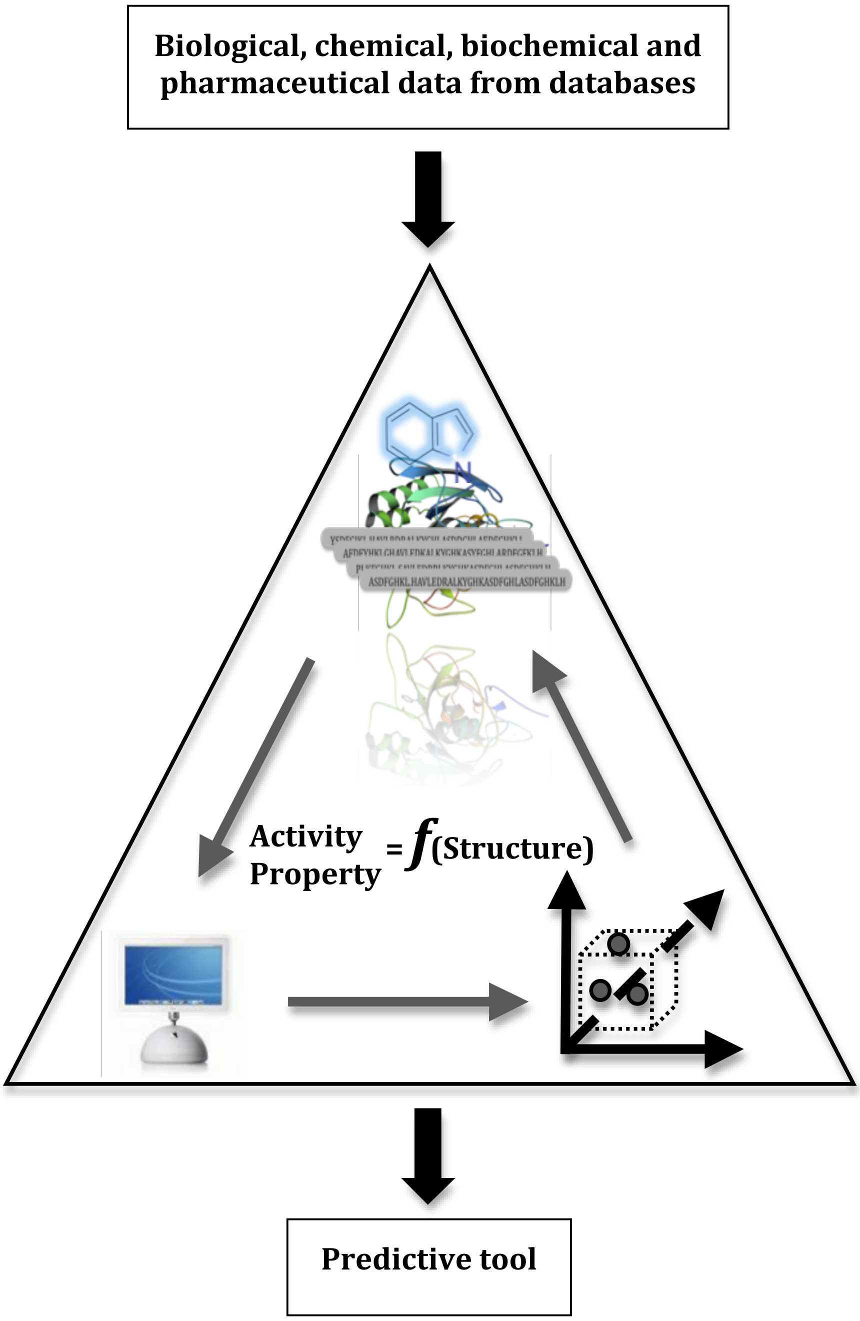

Biological, chemical, biochemical and pharmaceutical data from databases
Activity
Property
= f(Structure)
Predictive tool

Figure 2.2. Scheme of QSAR framework.

2.5.1 Topological autocorrelation vectors

The binding of a ligand to a target depends on the shape of the ligand and on a variety of factors such as molecular electrostatic potential, polarizability, hydrophobicity and lipophobicity. Similarly, the conformational stability of proteins depends on different intramolecular interactions such as hydrophobic, electrostatic, van der Waals, and hydrogen bond that are ruled by the amino acid sequence. Therefore, in a QSAR study the strategy for encoding molecular information, in some way, either explicitly or implicitly, should account for these physicochemical effects. Furthermore, datasets usually include molecules of with different numbers of atoms and/or proteins of different lengths, so the structural encoding schemes must allow comparing such molecules.

There are more than 3000 molecular descriptors reported for QSAR analysis, which include molecular properties (ClogP, HOMO and LUMO energy, etc), substructural fragments, molecular fingerprints, sophisticate topological and tridimensional indexes and so on [43]. Among them, autocorrelation vectors are very popular and easy to implement topological descriptors widely used in biological QSAR. Autocorrelation vectors can encode variable length structures into fixed-length information matrices having several useful properties. First, a substantial reduction in data can be achieved by limiting the topological distance, *l*. Second, the autocorrelation coefficients are independent of the original atom numbering, so they are canonical. And third, the length of the correlation vector is independent of the size of the molecule or protein sequence [44].

2.5.1.1 Topological autocorrelation vectors for ligands

For the autocorrelation vectors of ligands, H-depleted molecular structure is represented as a 2D graph and physico-chemical properties of atoms (i.e. atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities) as real values assigned to the graph vertices.

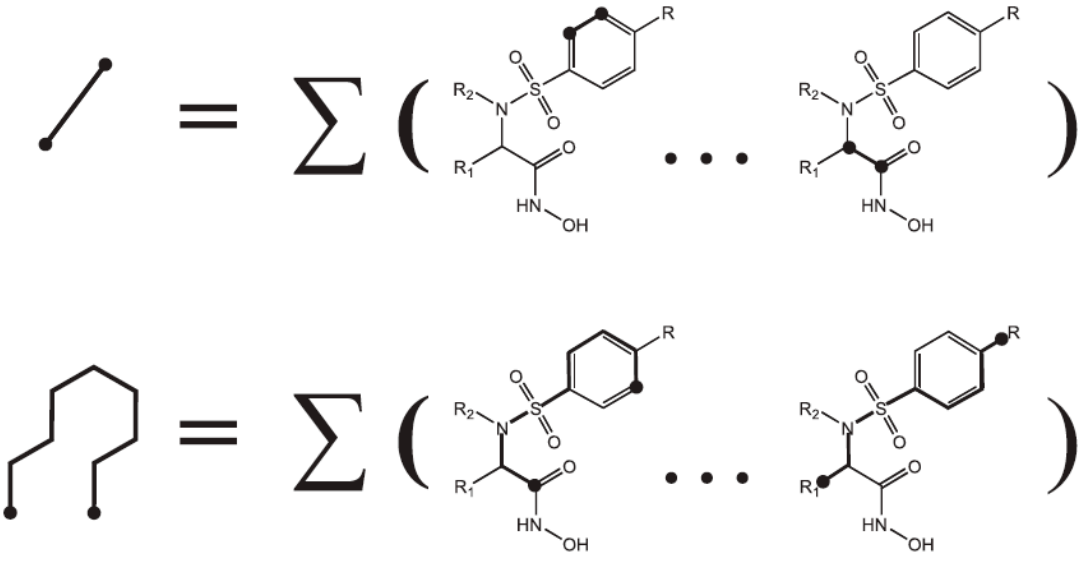


Figure 2.3. Representation of 2D autocorrelation terms at topological distances 1 and 8 in generic N-hydroxy-2-[(phenylsulfonyl) amino]acetamide derivative.

These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in the molecular graph (Figure 2.3). Broto-Moreau's autocorrelation vectors were employed for encoding the topological structure of the kinase inhibitors.

Broto-Moreau's autocorrelation coefficient [45] is defined below:

$$ATSlp_k = \sum_i \delta_{ij} p_{ki} p_{kj} \tag{2.1}$$

where *ATSlp*$_k$ is Broto-Moreau's autocorrelation coefficient at spatial lag *l*; p_{ki} and p_{kj} are the values of property *k* of atom *i* and *j*, respectively, and $\delta(l,\ d_{ij})$ is a delta function defined as:

$$\delta(l, d_{ij}) = \begin{Bmatrix} 1\ if\, d_{ij} = l \\ 0\ if\, d_{ij} \neq l \end{Bmatrix} \tag{2.2}$$

where d_{ij} is the topological distance or spatial lag between atoms *i* and *j*.

Dragon computer software [46] was used for calculating the 2D autocorrelation vectors at spatial lags ranging from 1 to 8 and weighted by 3 atomic properties: atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities, thus a total of 24 (8×3) 2D autocorrelation vectors were computed.

2.5.1.2 Amino acid Sequence Autocorrelation Vectors (AASA)

Autocorrelation vector formalism can be easily extended to amino-acid sequences considering protein primary structure as a linear graph with nodes formed by amino-acid residues. We introduced the *AASA* vectors for modeling the functional variations upon mutation of the ghrelin receptor [37] and the conformational stability of human lysozyme [30], gene V protein [31] and chymotrypsin inhibitor 2 mutants [47]. The calculated autocorrelation vectors encode information concerning whole protein sequence. Particularly, *AASA* vectors of lag *l* are calculated as follows:

$$AASAlp_k = \frac{1}{L}\sum_i \delta_{ij} p_{ki} p_{kj} \qquad (2.3)$$

where $AASAlp_k$ is the *AASA* at spatial lag *l* weighted by the p_k property; *L* is the number of elements in the sum; p_{ki} and p_{kj} are the values of property *k* of amino acids *i* and *j* in the sequence respectively and $\delta(l, d_{ij})$ is the delta function in Eq. 2.2.

For example, if we consider the decapeptide ASTCGFHCSD, *AASA* vectors at spatial lag 1 and 5 are calculated as follows:

$$AASA1p_k = \frac{1}{9}(p_{kA}\cdot p_{kS} + p_{kS}\cdot p_{kT} + p_{kT}\cdot p_{kC} + p_{kC}\cdot p_{kG} + p_{kG}\cdot p_{kF} + p_{kF}\cdot p_{kH} + p_{kH}\cdot p_{kC} + p_{kC}\cdot p_{kS} + p_{kS}\cdot p_{kD}) \qquad (2.4)$$

$$AASA5p_k = \frac{1}{5}(p_{kA}\cdot p_{kF} + p_{kS}\cdot p_{kH} + p_{kT}\cdot p_{kC} + p_{kC}\cdot p_{kS} + p_{kG}\cdot p_{kD}) \qquad (2.5)$$

In a protein, autocorrelation analysis tests whether the value of a property at one residue is independent of the values of the property at neighboring residues. If dependence exists, the property is said to exhibit spatial autocorrelation. *AASA* vectors represent the degree of similarity between amino acid sequences.

To model conformational stability, 48 physicochemical and conformational amino acid/residues properties (Table 2.1 Appendix) from the AAindex database were used as weights for sequence residues [28]. The spatial lag, *l*, was ranging from 1 to 15. A matrix of 720 *AASA* vectors, 48 properties × 15 different lags, was generated with the autocorrelation vectors calculated for each target. *AASA* vectors were calculated by Protmetrics in-house software [48].

To model inhibition of kinases and proteases, seven physicochemical and conformational amino acid/residues properties (Table 2.2 Appendix) from the AAindex database [28] were used as weights for sequence residues. The spatial lag, *l*, was ranging from 1 to 5. A matrix of 35 *AASA* vectors, 7 properties × 5 different lags, was generated with the autocorrelation vectors calculated for each target.

2.5.2 Structural fragments features

2.5.2.1 Structural Fragments (SF) descriptors for ligands

SF were computed for ligands by counting 120 fragments in the chemical structures by Dragon computer software [46].

2.5.2.2 Sequence Structural Fragments (SSF) descriptors for protein sequences

SF descriptors were computed for the 20 amino acids by Dragon computer software [46] and relative amino-acid compositions were computed for the kinase and protease sequences by a Matlab [49] code. Afterwards, Sequence Structural Fragments (*SSF*) descriptors (120×1 row vector) were calculated in Matlab [49] for each kinase and protease as the matrix product of the amino-acid *SF* descriptors (120×20 matrix) by the relative amino-acid composition of kinases and proteases (20×1 column vector).

2.6. Machine Learning in Computational Biology

In computational biology, the existing experimental data is use to built predictive models to annotate related information yet experimentally measured [50]. Different mathematical models are used to predict the effect of structural features on the properties and functions of biomolecules. This knowledge has been employed to compute the characteristics of experimentally unmeasured elements of the same or related systems from novel bioactive compounds to new functional regions or fragments in protein and nucleic acid sequences. The implementation of machine learning in computational biology has grown up rapidly. These techniques have encountered successful applications in structure-function/property studies to predict protein secondary structure [51-53], protein subcellular location [33-34], enzyme active sites [54-56], solvent accessibility [57], protein-protein interactions [58,59], protein-DNA/RNA interactions [60-62], protein conformational stability [2, 2, 30, 31], ligand affinity [44, 63-65], etc. Artificial neural networks, support vector machines and random forest models have been among the most popular approaches to map protein properties and ligand affinities to structural features.

2.6.1 Support Vector Machines (SVMs)

SVM is a new machine learning method, which has been used for many kinds of pattern recognition problems. Since there are excellent introductions to SVMs [66] only the main idea of SVMs applied to pattern classification problems is stated here. Firstly, the input vectors are mapped into one feature space (possible with a higher dimension). Secondly, a hyperplane, which can separate two classes, is constructed within this feature space. Only relatively low-dimensional vectors in the input space and dot products in the feature space will evolve by a mapping function. SVM was designed to minimize structural risk whereas previous techniques were usually based on minimization of empirical risk. SVM is less vulnerable to the overfitting problem, so it can deal with a large number of features.

The mapping into the feature space is performed by a kernel function. There are several parameters in the SVM, including the kernel function and regularization parameter. The kernel function and its specific parameters, together with regularization parameter, cannot be set from the optimization problem but have to be externally tuned. These can be set by Vapnik-Chervonenkis bounds, crossvalidation, an independent optimization set, or Bayesian learning [67]. In this work, the Radial Basic Function (RBF) was used as kernel function. The toolbox used to implement the SVM with RBF kernel was libSVM for Matlab [49] by Chang and Lin [68] that can be downloaded from: http://www.csie.ntu.edu.tw/cjlin/libsvm/.

Since >300 autocorrelation vectors were available to model protein conformational stability, GA-based SVM (GA-SVM) algorithm was implemented for choosing the optimum subset of input training vectors and setting the two SVM parameters, regularization parameter and width of the RBF kernel. The optimization inside the GA framework was driven by crossvalidation.

In the case of kinase and protease inhibition feature selection step was not necessary because an easily manageable amount of 59 autocorrelation vectors were available.

2.6.2 Genetic Algorithm-optimized Support Vector Machines (GA-SVMs)

The application of machine learning for solving classification and function mapping problems in computational biological has vastly grown in the last years. However, it is difficult to choose the adequate descriptors for predictor training due to lack of absolute rules that govern this choice. Evolutionary algorithms and specifically Genetic Algorithm have been used for variable selection problems [30, 31]. Since 302 *AASA* vectors were available for modelling and only a subset of them is statistically significant in terms of correlation with the mutants stability, it was needed implementing an optimal model by variable selection.

In the case of protein conformational stability, GA was applied at the same time for selection of the optimum subset of variables and also to the optimization of regularization parameter and width of an RBF kernel, according to Fröhlich et al. [67]. We can simply concatenate a representation of the parameter to a chromosome representing subset of variables used for SVM training. That means we are trying to select an optimal feature subset and a regularization parameter at the same time. This is reasonable because the choice of the parameter is influenced by the feature subset taken into account and vice versa. Usually it is not necessary to consider any arbitrary value except certain discrete values with the form: $n\times10^{k}$, where n=1…9 and k=-4…4. So, these values can be calculated randomly generating *n* and *k* values as integers between (1…9) and (-4…4), respectively. In a similar way we used GA to optimize the width of an RBF kernel. Then, our chromosome was concatenated with another gene with discrete values in the interval (0.001-90 000) for encoding the regularization parameter and the width of the RBF kernel.

A Five-Fold-Out (FFO) crossvalidation assessed model quality throughout the GA search. Five data subsets were created, four subsets are generated in the crossvalidation process for training the SVM and another subset is then predicted. This process is repeated until all subsets have been predicted. A “venetian-blind” method was used for creating the data subsets. In the first place, dataset is ordered according to the dependent variable and in the second step the cases are added consecutively to each subset, in such a way that they become representative samples of the whole dataset. In order to avoid overestimation of the model’s predictive power, similar mutants were kept in the same set during crossvalidation, even when they reported under different experimental Temperature and pH values. The GA routine minimized the regression mean squared error of FFO (MSE_{FFO}) crossvalidation experiment.

Afterwards, the same subset of optimum variables selected by the regression GA-SVM was used for training a SVM classifier. Nevertheless regularization parameter and width of RBF kernel for the SVM binary classifier were set by a bidimensional grid search around optimum GA-selected parameters, which minimizes the percent of misclassifications of FFO crossvalidation (MC_{FFO}).

A version of the GA previously reported by us [37] was applied here to SVM hyperparameters optimization and feature selection for modelling conformational stability. GlibSVM [69] toolbox for Matlab was programmed within Matlab environment [49] using Genetic Algorithm [70] and libSVM Toolboxes [68].

2.6.3 Model’s validation

The quality of the regression SVM models was evaluated by the squared correlation coefficient of FFO crossvalidation (R^2_{FFO}) and the root MSE_{FFO} ($RMSE_{FFO}$) and also calculated classification statistics of test set.

The efficiency of the SVM predictor for the classification problem was accomplished using the set of statistics listed below.

The overall accuracy is

$$Q2 = \frac{p}{N} \tag{2.6}$$

where *p* is the total number of correct predicted instance and *N* is the total number of instances.

The correlation coefficient *Cr* is defined as follow:

$$Cr(s) = \frac{\left[p(s)n(s) - u(s)o(s)\right]}{D} \tag{2.7}$$

where *D* is the normalization factor

$$D = \left[\left(p(s)+u(s)\right)\left(p(s)+o(s)\right)\left(n(s)+u(s)\right)\left(n(s)+o(s)\right)\right]^{1/2} \tag{2.8}$$

for each class s (+ and – for positive and negative instance); *p(s)* and *n(s)* are the number of correct predictions and correctly rejected assignments, respectively, and *u(s)* and *o(s)* are the number of under- and over-predictions.

The coverage for each discriminant structure *s* is evaluated as

$$Q_S = \frac{p(s)}{p(s)+u(s)} \tag{2.9}$$

where *p(s)* and *u(s)* are the same as in Eq. 2.8

The accuracy for s is computed as

$$P_S = \frac{p(s)}{p(s) + o(s)} \qquad (2.10)$$

where *p(s)* and *u(s)* are the same as in Eq. 2.8.

F-score known as the harmonic mean of sensitivity and positive precision given as follows:

$$F = 2 \times \frac{\left(Q(+) \times P(+)\right)}{\left(Q(+) + P(+)\right)} \qquad (2.11)$$

2.7. References

1. Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34**:D204–D206.

2. Capriotti E, Fariselli P, Casadio R: **A neural-network-based method for predicting protein stability changes upon single mutations**. *Bioinformatics* 2004, **20**:63–68.

3. Ahmad S, Kitajima K, Selvaraj S, Kubodera H, Sunada S, An J-H, Sarai A: **Protein-Ligand Interactions: ProLINT Database and QSAR Analysis.** *Genome Inform* 2003, **14**:537–538.

4. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker W C, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: **The Universal Protein Resource (UniProt): an expanding universe of protein information**. *Nucleic Acids Res* 2006, **34**:D187–D191.

5. **Instant JChem ChemAxon, version 2.1.1**, ChemAxon Ltd 2007, Budapest, Hungary.

6. Roman-Roldan R, Bernaola-Galvan P, Oliver JL: **Application of information theory to DNA sequence analysis: a review**. *Patt Recogn* 1996, **29**:1187–1194.

7. Nady A: **Recent investigations into global characteristics of long DNA sequences**. *Indian J Biochem Bioph* 1994, **31**:149–155.

8. Tio P: **Spatial representation of symbolic sequences through iterative function systems**. *IEEE T Syst Man Cy A* 1999, **29**:386–393.

9. Jeffrey HJ: **Chaos game representation of gene structure**. *Nucleic Acid Res* 1990, **18**: 2163–2170.

10. Forte B, Mendivil F, Vrscay ER: **Chaos games for iterated function systems with grey level maps**. *SIAM J Math Anal* 1998, **29**:878–890

11. Fiser A, Tusnády GE, Simon I: **Chaos game representation of protein structures**. *J Mol Graph* 1994, **12**: 302–304.

12. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B: **Genomic signature: characterization and classification of species assessed by chaos game representation of sequences**. *Mol Biol Evol* 1999, **16**: 1391–1399.

13. Enright MC, Knox K, Griffiths D, Crook DWM, Spratt BG: **Multilocus sequence typing of Streptococcus pneumoniae directly from cerebrospinal fluid**. *Eur J Clin Microbiol Infect Dis* 2001, **19**:627–630.

14. Roy A, Raychaudhury C, Nandy A: **Novel techniques of graphical representation and analysis of DNA sequences – a review**. *J Biosci* 1998, **23**:55–71.

15. Vinga S, Almeida J: **Alignment-free sequence comparison**. *Bioinformatics* 2003; **19**:513–523.

16. Almeida JS, Vinga S: **Universal sequence map (USM) of arbitrary discrete sequences**. *BMC Bioinformatics* 2002, **3**:6–17.

17. Zu-Guo Y, Vo A, Ka-Sing L: **Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses**. *J Theor Biol* 2004, **226**:341–348.

18. Aguero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Dıaz Y: **Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L**. *FEBS Lett* 2006, **580**: 723–730.

19. Aguero-Chapin G, Gonzlez-Diaz H, Riva G, Rodriguez E, Sanchez-Rodriguez A, Podda G, Vazquez-Padron RI: **MMM-QSAR Recognition of Ribonucleases without Alignment: Comparison with an HMM Model and Isolation from Schizosaccharomyces pombe, Prediction, and Experimental Assay of a New Sequence**. *J Chem Inf Model* 2008, **48**:434–448.

20. Dea-Ayuela MA, Pérez-Castillo Y, Meneses-Marcela A, Ubeira FM, Bolas-Fernández F, Kuo-Chen C, González-Díaz H: **HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence**. *Bioorg Med Chem* 2008, **16**:7770–7776.

21. Bai F, Wang T: **On Graphical and Numerical Representation of Protein Sequences**. *J Biomol Struct Dyn* 2006, **23**:537–545.

22. Fernández M, Caballero J, Fernández L, Abreu JI, Acosta G: **Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines**. *Proteins* 2008, **70**:167–175.

23. Fernández M, Fernández L, Abreu JI, Garriga M: **Classification of voltage–gated K(+) ion channels from 3D pseudo–folding graph representation of protein sequences using genetic algorithm–optimized support vector machines**. *J Mol Graph Model* 2008, **26**:1306–1314.

24. Randić M, Zupan J, Vikić-Topić D: **On representation of proteins by star-like graphs**. *J Mol Graph Model* 2007, **26**:290–205.

25. Munteanu CR, Gonzalez-Diaz, H, Borges F, Lopes de Magalhnes A: **Natural/random protein classification models based on star network topological indices**. *J Theor Biol* 2008, **254**:775–783.

26. Wolfram, S: **Cellular automata as models of complexity**. *Nature* 1984, **31**: 419–424.

27. Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou, K-C: **Using cellular automata to generate image representation for biological sequences**. *Amino Acids* 2005, **28**:29–35.

28. (a) Nakai K, Kidera A, Kanehisa M: **Cluster analysis of amino acid indices for prediction of protein structure and function**. *Protein Eng* 1988, **2**:93–100. (b) Tomii K, Kanehisa M: **Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins**. *Protein Eng* 1996, 9:27–36. (c) Kawashima S, Kanehisa M: **AAindex: amino acid index database**. *Nucleic Acids Res* 2000, 28:374.

29. Siebert JK: **Modeling Protein Functional Properties from Amino Acid Composition**. *J Agric Food Chem* 2003, **51**:7792–7797.

30. Caballero J, Fernández L, Abreu JI, Fernández M: **Amino Acid Sequence Autocorrelation Vectors and Ensembles of Bayesian-Regularized Genetic Neural Networks for Prediction of Conformational Stability of Human Lysozyme Mutants**. *J Chem Inf Model* 2006, **46**:1255–1268.

31. Fernández L, Caballero J, Abreu JI, Fernández M: **Amino Acid Sequence Autocorrelation Vectors And Bayesian-Regularized Genetic Neural Networks For Modeling Protein Conformational Stability: Gene V Protein Mutants**. *Proteins* 2007, **67**:834–853.

32. Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA: **Protein Linear Indices of the 'Macromolecular Pseudograph α-Carbon Atom Adjacency Matrix' in Bioinformatics. Part 1: Prediction of Protein Stability Effects of a Complete Set of Alanine Substitutions in Arc Represor**. *Bioorg Med Chem* 2005, 13:3003–1305.

33. Chou KC: **Prediction of protein cellular attributes using pseudoamino-acid-composition**. *Proteins* 2001, **43**:246–255 (Erratum: ibid. (2001) **44**: 60).

34. Chen C, Chen L, Zou X, Cai P: **Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine**. *Protein Pept Lett* 2009, **16**:27–31.

35. Ramos de Armas R, González-Díaz H, Molina R, Uriarte E: **Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants**. *Proteins* 2004, **56**:715–723.

36. González-Díaz H, Vilar S, Santana L, Uriarte E: **Medicinal chemistry and bioinformatics–current trends in drugs discovery with networks topological indices**. *Curr Top Med Chem* 2007, **7**:1015–1029.

37. Caballero J, Fernández L, Garriga M, Abreu JI, Collina S, Fernández M: **Proteometric Study of Ghrelin Receptor Function Variations upon Mutations using Amino Acid Sequence Autocorrelation Vectors and**

Genetic Algorithm-Based Least Square Support Vector Machines. *J Mol Graph Model* 2007, **26**:166-78.

38. Sheinerman FB, Giraud E, Laoui A: **High Affinity Targets of Protein Kinase Inhibitors Have Similar Residues at the Positions Energetically Important for Binding**. *J Mol Biol* 2005, **352**:1134–1156.

39. Kurup A, Garg R, Hansch C: **Comparative QSAR Study of Tyrosine Kinase Inhibitors**. *Chem Rev* 2001, **101**:2573–2600.

40. Verma RP, Hansch C: **Matrix metalloproteinases (MMPs): Chemical–biological functions and (Q)SARs**. *Bioorg Med Chem* 2007, **15**:2223–2268.

41. Strömbergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JES, Komorowski J, Hvidsten TR: **Generalized Modeling of Enzyme–Ligand Interactions Using Proteochemometrics and Local Protein Substructure**. *Proteins* 2006, 65, 568–579.

42. Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, Wikberg JES: **Development of proteo-chemometrics: A novel technology of use for analysis of drug-receptor interactions**. *Biochem Biophys Acta* 2001, **1525**:180–190.

43. Todeschini R, Consonni V: *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH; 2000.

44. Bauknecht H, Zell A, Bayer H, Levi P, Wagener M, Sadowski J, Gasteiger J: **Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists**. *J Chem Inf Comput Sci* 1996, **36**:1205–1213.

45. Moreau G, Broto P: **Autocorrelation of a topological structure: A new molecular descriptor**. *Nouv J Chim* 1980, **4**:359–360.

46. DRAGON, version 3.0; Milano Chemometrics: Milan, Italy, 2003.

47. Fernández M, Abreu JI, Caballero J, Garriga M, Fernández L: **Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks**. *Mol Simulat* 2007, **13**:1045–1056.

48. Fernandez M, Abreu JI: Protmetrics, Molecular Modeling Group, University of Matanzas, Cuba, 2006.

49. MATLAB, version 7.0; The Mathworks Inc.: Natick, MA, 2006.

50. Dodson EJ: **Computational biology: Protein predictions**. Nature 2007, **450:**176–177.

51. Ward JJ, McGuffin LJ, Buxton BF, Jones DT: **Secondary structure prediction with support vector machines**. *Bioinformatics* 2003, **19**:1650–1655.

52. Vullo A, Bortolami O, Pollastri G, Tosatto SCE: **Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines**. *Nucleic Acids Res* 2006, **34**:W164–W168.

53. Zhang Q, Yoon S, Welsh WJ: **Improved method for predicting ß-turn using support vector machines.** *Bioinformatics* 2005, **21**:2370–2374

54. Blom N, Gammeltoft S, Brunak S: **Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites**. *J Mol Bio* 1999, **294**:1351–1362.

55. Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I: **NetPhosBac - a predictor for Ser/Thr phosphorylation sites in bacterial proteins**. *Proteomics* 2009, **9**:116–125.

56. Bhasin M, Raghava GPS: **Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition**. *J Biol Chem* 2004, **279**:23262–23266.

57. Ahmad S, Gromiha MM, Sarai A: **RVP-net: online prediction of real valued accessible surface area of proteins from single sequences**. *Bioinformatics* 2003, **19**:1849–1851.

58. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, Luo X, Golshani A: **PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.** *BMC Bioinformatics* 2006, **7**:365–380.

59. Chen X-W, Liu M: **Prediction of protein–protein interactions using random decision forest framework**. *Bioinformatics* 2005, **21**:4394–4400.

60. Ahmad S, Gromiha MM, Sarai A: **Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information**. *Bioinformatics* 2004, **20**:477–486.

61. Ahmad S, Sarai A: **PSSM-based prediction of DNA binding sites in proteins**. *BMC Bioinformatics* 2005, 6:33–38.

62. Spriggs RV, Murakami Y, Nakamura H, Jones S: **Protein function annotation from sequence: prediction of residues interacting with RNA**. *Bioinformatics* 2009, **25**:1492–1507.

63. Fernández M, Tundidor-Camba A, Caballero J: **Modeling of Cyclin-Dependent Kinase Inhibition by 1H-pyrazolo [3,4-d] pyrimidine Derivatives using Artificial Neural Networks Ensembles**. *J Chem Inf Model* 2005, **45**:1884–1895.

64. González MP, Caballero J, Helguera AM, Garriga M, González G, Fernández M: **2D autocorrelation modeling of the inhibitory activity of cytokinin-derived cyclindependent kinase inhibitors**. *Bull Math Biol* 2006, **68**:735–751.

65. Caballero J, Fernandez M, Saavedra M, Gonzalez-Nilo FD: **2D Autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido[2,3-d]pyrimidine derivatives**. *Bioorg Med Chem* 2008, **16**:810–821.

66. a) Cortes C, Vapnik V: **Support-Vector Networks**. *Mach Learn* 1995, **20**: 273–297. b) Burges CJC: **A tutorial on support vector machines for pattern recognition**. *Data Min Knowl Discov* 1998, **2**:1-47. c) Vapnik V: Statistical Learning Theory. New York: Wiley; 1998.

67. Fröhlich H, Chapelle O, Schölkopf B: **Feature Selection for Support Vector Machines by Means of Genetic Algorithms**. *In Proceedings of the 15th IEEE Int. Conf. on Tools with AI.* 2003, 142–148.

68. Chih-Chung C, Chih-Jen L. LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm; 2001.

69. Fernandez M. GlibSVM toolbox for Matlab version 1.0, Molecular Modeling Group, University of Matanzas; 2007.

70. The MathWorks Inc. **Genetic algorithm and direct search toolbox user's guide for use with MATLAB**, Massachusetts: The Mathworks Inc., 2004.

CHAPTER 3. MODELING OF PROTEIN CONFORMATIONAL STABILITY

3.1. Introduction

This chapter presents the modeling of protein conformational stability from the amino acid sequences. Conformational stability was predicted by extending the concept of structural autocorrelation vectors [1-6] in molecules to protein primary structure. Protein sequence was encoded by means of Amino Acid Sequence Autocorrelation (*AASA*) vectors weighted by 48 physicochemical, energetic, and conformational amino acid/residue properties extracted from the AAindex amino acid database [7]. Robust classification and regression predictors of the conformational stability of protein mutants were trained using sequence information from a dataset of 1383 non-redundant mutants from Prothem database [8]. In addition to internal crossvalidation, the stability of a test set of non-redundant 222 single, 277 double and 144 multiple-point mutations were also classified. A large set of 720 *AASA* vectors was calculated on the mutant sequences. After eliminating intercorrelated vectors, a total of 302 autocorrelation vectors were available for building the models. Then, function mapping of conformational stability was accomplished by training regression genetic algorithm-optimized support vector machines (GA-SVM) that minimize the error of Five- Fold Out (FFO) crossvalidation. In the GA-SVM framework, optimum subset of training *AASA* inputs and SVM parameter values were set using genetic algorithm (GA) rules. The optimum SVM was further trained to classify stable and unstable mutants but the regularization parameter and width of the radial basis kernel (RBF) kernel were set by grid search. Afterwards, normalized Temperature and pH values of the $\Delta\Delta G$ experimental measures were added to the SVM in order to improve predictor performance and prediction accuracies for different mutations types were evaluated.

3.2. Results

3.2.1 Function mapping of mutants conformational stability

Firstly, GA-SVM approach was applied to build optimum nonlinear regression models of protein conformational stability using nonlinear RBF kernel inside the SVM framework. Subspaces in the dataset were searched varying number of training variables from 5 to 30. From one generation to another GA minimized the MSE_{FFO} with *FFO* crossvalidation subsets selected according to "venetian blind" method. In addition to selecting the optimum input subset, the GA optimized the kernel regularization parameter and the width of an RBF kernel.

Table 3.1. Crossvalidation statistics of the SVM model for prediction of protein mutant *ΔΔG* real values. SVM regularization parameter was 1 and width of the RBF kernel was 0.071.

Regression SVM inputs	R^2_{FFO}	$RMSE_{FFO}$
$AASA11N_m$, $AASA8P$, $AASA7P_B$ $AASA7G_{hN}$, $AASA10H_t$, $AASA14f$, $AASA12\Delta G_C$, $AASA15\Delta ASA$, $AASA15\Delta Cp_h$, $AASA14\Delta Cp_h$	0.42	0.139
$AASA11N_m$, $AASA8P$, $AASA7P_B$ $AASA7G_{hN}$, $AASA10H_t$, $AASA14f$, $AASA12DG_C$, $AASA15\Delta ASA$, $AASA15\Delta Cp_h$, $AASA14\Delta Cp_h$, Temperature, pH	0.45	0.136

R^2_{FFO} and $RMSE_{FFO}$ are the square correlation coefficient and the root mean square error of five-fold-out (FFO) crossvalidation.

Table 3.1 shows inputs, parameters and statistical quantities for data fitting and crossvalidation experiment of the optimum SVM predictor (input features names appear in Table 3.1 Appendix). Optimum regularization parameter and width of the RBF kernel were 1 and 0.071, respectively. The optimum autocorrelation vector subset only contains two significant pair correlations (R^2>0.7): *AASA14f* vs. *AASA15ASA$_N$* and *AASA15ΔCp$_h$* vs. *AASA14ΔCp$_h$*. Despite this little intercorrelation, the good correlation in FFO crossvalidation reflects that relevant structural information is brought into the model by each *AASA* descriptor.

Figure 3.1A depicts plot of calculated vs. experimental *ΔΔG* values in crossvalidation experiment according to the optimum SVM model with 10 *AASA* vectors with correlation coefficient of 0.65. In order to increase the predictive accuracy of the model, Temperature and pH values of the experimental determinations were added as extra inputs to the regression SVM plotted in Figure 3.1B. The correlation coefficient was increased up to 0.67 representing about 50% of explained crossvalidation data variance.

It is interesting to evaluate the performance of the predictor regarding the nature of the mutations. Mutations were classified according to the physico-chemical properties of the substituted and new residues. Figure 3.2 depicts plots of calculated FFO vs. experimental *ΔΔG* values for mutants according to mutation types. The lowest predictions were yielded for charged/charged, polar/charged and apolar/charged. The specific effects of residue substitutions on the real *ΔΔG* values are better predicted for polar/polar, and polar/apolar mutations with crossvalidation accuracy over 50%. In addition, the accuracies of the prediction according to the type of secondary structure of the mutation site were investigated. The type of secondary structure was assigned to each mutation from the database Protherm [8], in which residues are classified in four different secondary structures: helix, sheet, turn and coil. The lowest correlations were found for mutations allocated at helix and turn structures (Figure 3.1 Appendix).

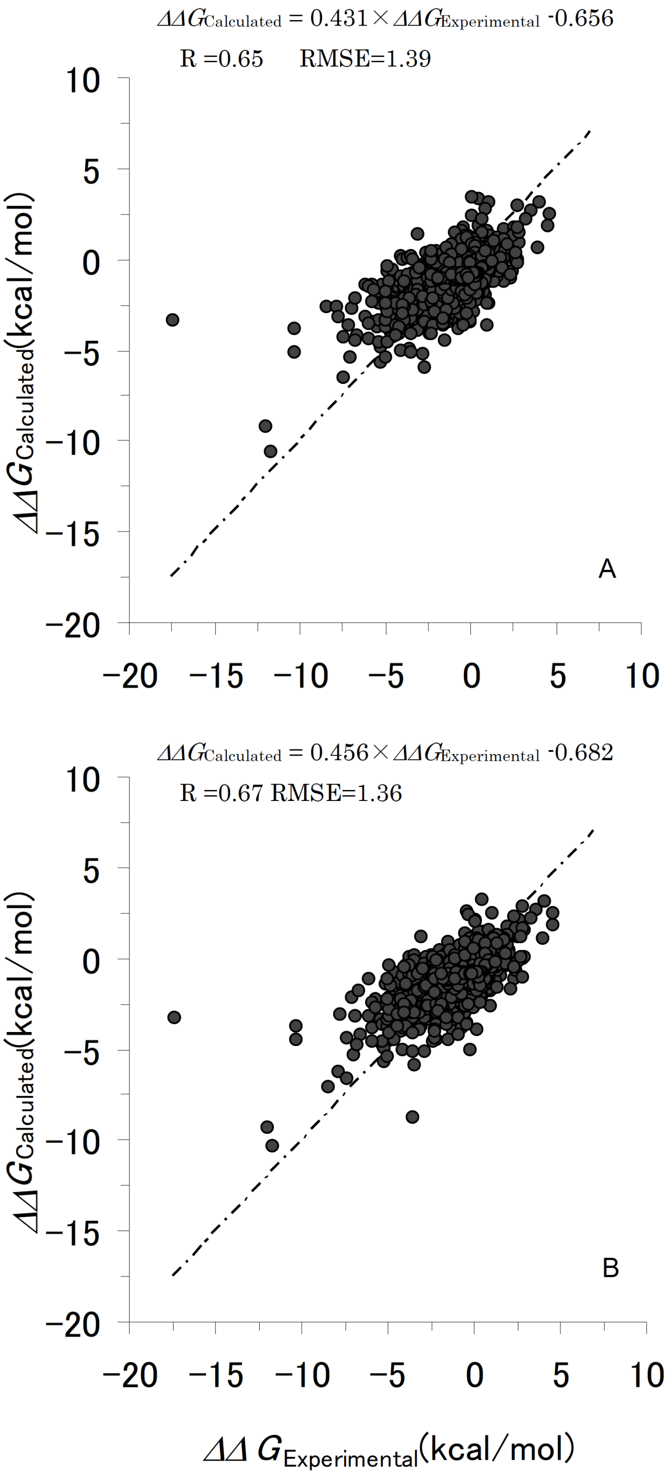


Figure 3.1. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change (*ΔΔG*) of protein mutants according to regression SVM models without including experimental condition data (A) and including experimental condition data (B) as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

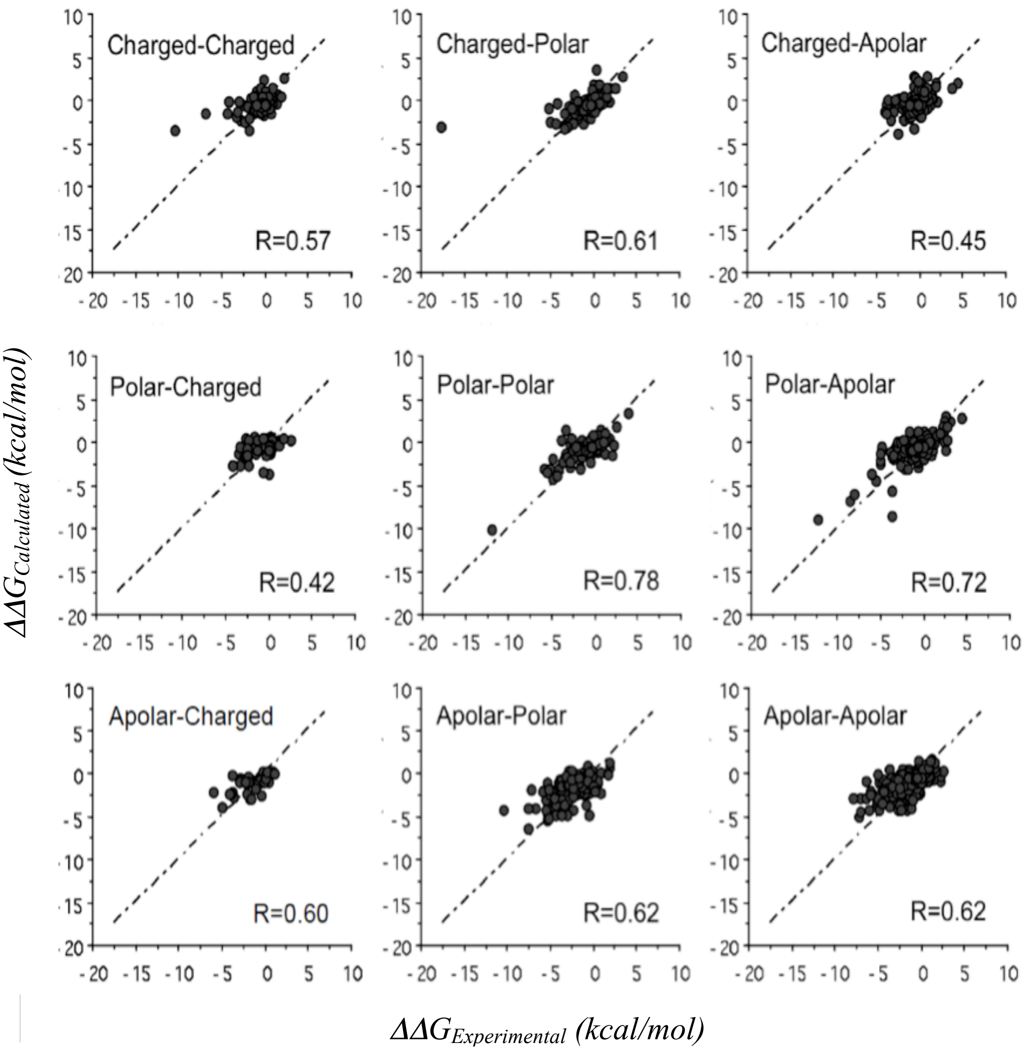


Figure 3.2. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change (*ΔΔG*) of protein mutants for each mutation type according to regression SVM models including experimental condition data as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

Similarly, it was carried out an analysis of the regression accuracy taking into account the accessible surface area (ASA) values of the mutation sites. According to Protherm database, mutations were grouped as buried (ASA<20%), partially buried (50%<ASA<20%) and exposed (ASA>50%). The correlation coefficients decrease with the decrements of ASA values of the mutated site (Figure 3.2 Appendix). The lowest regression accuracies correspond to exposed mutations allocated at the proteins surface and the highest correlation value to mutations in buried sites at protein core.

Table 3.2. Crossvalidation statistics of the SVM model for the classification of protein mutant *ΔΔG* signs. SVM regularization parameter was 7 and width of the RBF kernel was 0.167.

SVM inputs	*Q2*	*P(+)*	*P(-)*	*Q(+)*	*Q(-)*	*C*
$AASA11N_m$, $AASA8P$, $AASA7P_B$ $AASA7G_{hN}$, $AASA10H_t$, $AASA14f$,$AASA12\Delta G_C$, $AASA15ASA_N$, $AASA15\Delta Cp_h$, $AASA14\Delta Cp_h$	0.78	0.59	0.87	0.68	0.82	0.48
$AASA11N_m$, $AASA8P$, $AASA7P_B$ $AASA7G_{hN}$, $AASA10H_t$, $AASA14f$,$AASA12\Delta G_C$, $AASA15ASA_N$, $AASA15\Delta Cp_h$, $AASA14\Delta Cp_h$, Temperature, pH	0.77	0.57	0.88	0.71	0.80	0.48

+ and - : the indexes were evaluated for positive and negative *ΔΔG* signs. *Q2* is the number of correct predictions/number of examples; *P(s)* is the number of correct prediction for class s/all prediction made for s; *Q(s)* is the number of correct prediction for class s/observed in class s; *Cr* is Matthews's correlation coefficient. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction and *P(+)* and *P(-)* are precision scores.

3.2.2 Classification of mutants conformational stability

In addition to the regression SVM model, a binary classifier for the recognition of stable and unstable mutants was built. The optimum *AASA* vector subset was used for training the binary SVM classifier and minimizing MC_{FFO} in a grid search set the SVM hyperparameters. Optimum values of regularization parameter and RBF kernel width were 7 and 0.167 respectively, which yield training and crossvalidation results in Table 3.2. As can be observed, the binary SVM trained with optimum *AASA* subset has overall FFO crossvalidation accuracy of 78% and correlation coefficient *Cr*=0.48. It is noteworthy that the crossvalidation statistics for recognizing stable mutants *Q(+)*=0.68 and unstable mutants *Q(-)*=0.77 are in the range of the overall accuracy achieved. This result is quite interesting since the predictor only used sequence information encoded in 10 *AASA* vectors. Afterward, Temperature and pH normalized values were passed into the binary SVM as extra inputs and *Q2*, *Q(+)* and *Q(-)* values increased up to 0.77, 0.71 and 0.80 respectively, with a correlation coefficient *Cr*=0.48. The binary classifier yielded the lowest accuracy for mutations of charged residues by other charged residues (Table 3.2 Appendix).

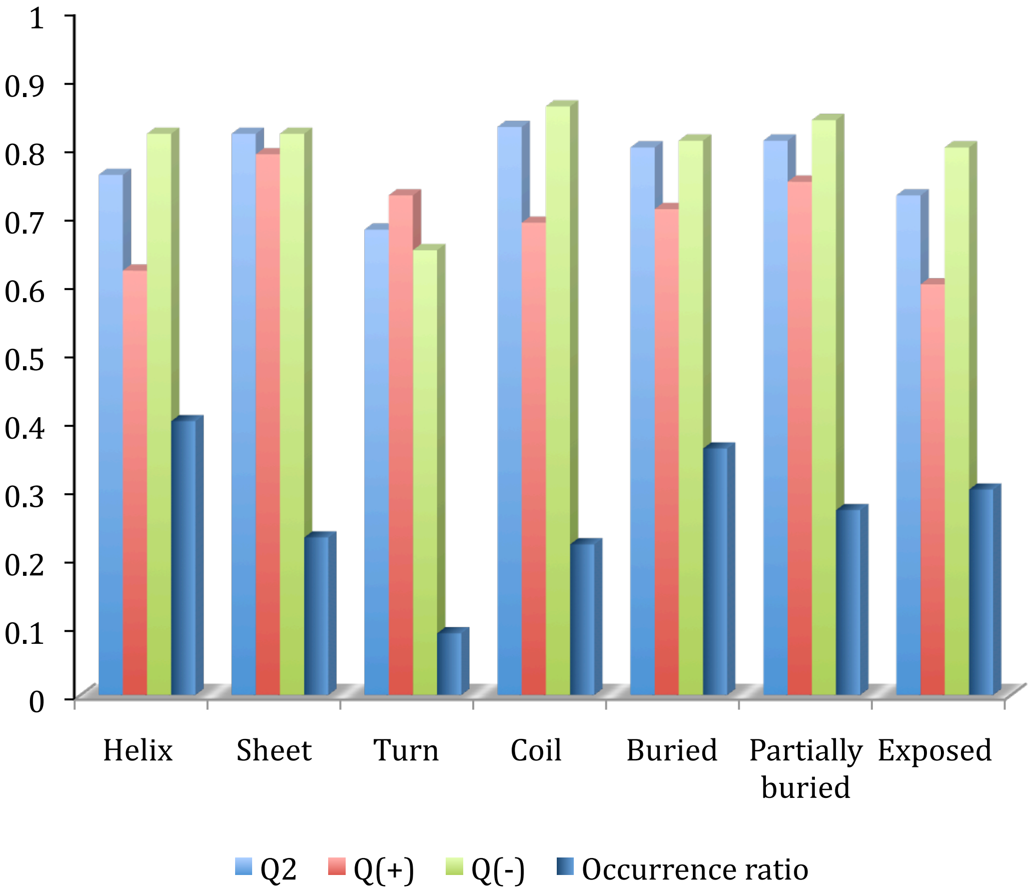


Figure 3.3. Crossvalidation classification accuracies of the optimum SVM model for the *ΔΔG* signs upon mutations according to secondary structure allocation and accessible surface area of the mutated residue. + and - : the indexes were evaluated for positive and negative *ΔΔG* signs. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction.

Furthermore, the accuracy of the classifier according to the type of secondary structure found in the mutation site was analyzed. Likewise, the regression model in Table 3.3 shows that, the lower accuracies were for mutations allocated at helix and coil structures. However, overall classification accuracies, about or higher than 70%, were observed for all types of mutations. Figure 3.3 shows the classification accuracies for the ASA values of the mutation sites. The overall accuracies for buried and partially buried mutations were about 80% and again the classification accuracy for mutations in the protein surface was lower around 70%. These results support the fact that protein properties, which depend on interactions at the protein surface, are more difficult to predict.

3.2.3 Classification of new mutants

Besides crossvalidation experiment, the sign of $\Delta\Delta G$ values was predicted for a test set with new single point mutants in Protherm database [8], all double and multiple mutations on this database. Prediction of real $\Delta\Delta G$ values was inaccurate. Results of stability classification of mutations in the test set are depicted in Figure 3.4. As it can be observed, the performance of the predictor on the single point mutation test set was poor. A lower overall accuracy about 51% was yielded with an adequate recognition of about a 66% of the stable mutations and low 45% of the unstable ones. Besides, the double point mutation test set showed $Q2$ = 0.50 with $Q(+)$ = 0.70 and $Q(-)$ = 0.34. Despite the discrete results, the predictor is able to account for single point mutation effects and to generalize them in some extent to double point mutations. It should be pointed out that unbalance classification results had been previously reported by Cappriotti et al. [8, 9] in crossvalidation experiments.

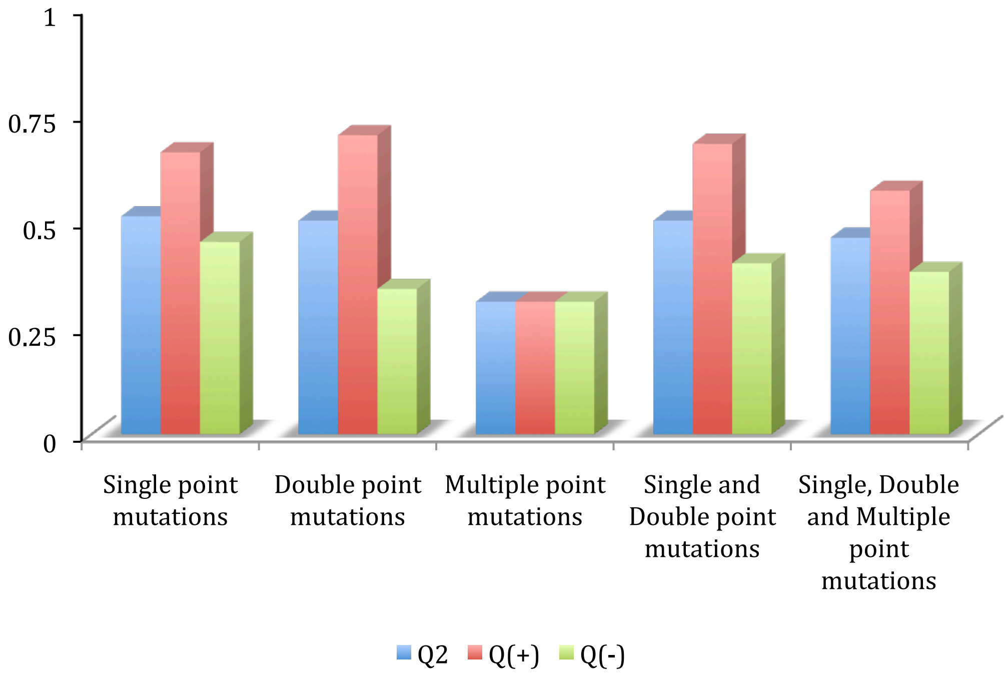


Figure 3.4. Test set classification accuracies according to SVM model for the classification of protein mutant *ΔΔG* signs. SVM regularization parameter was 7 and width of the RBF kernel was 0.167. + and - : the indexes were evaluated for positive and negative *ΔΔG* signs. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction.

Multiple mutations exhibited the lowest prediction results with low accuracies around 30%. On the contrary, the accuracies of single and double mutant test set were about 50%, 68% and 40% for all mutants, stable and unstable mutations. Meanwhile, these statistics for the whole test set (single, double and multiple mutants) were low with values of 46%, 57% and 38%. When considering only single and double point mutant, the test set overall accuracies were about 50% and for recognition of the stable single and double point mutants the classifier exhibited a higher accuracy about 70%. Finally, Figure 3.5 depicts classification accuracies for the single point mutants in the test set according to the secondary structure allocation and the ASA of the mutation site.

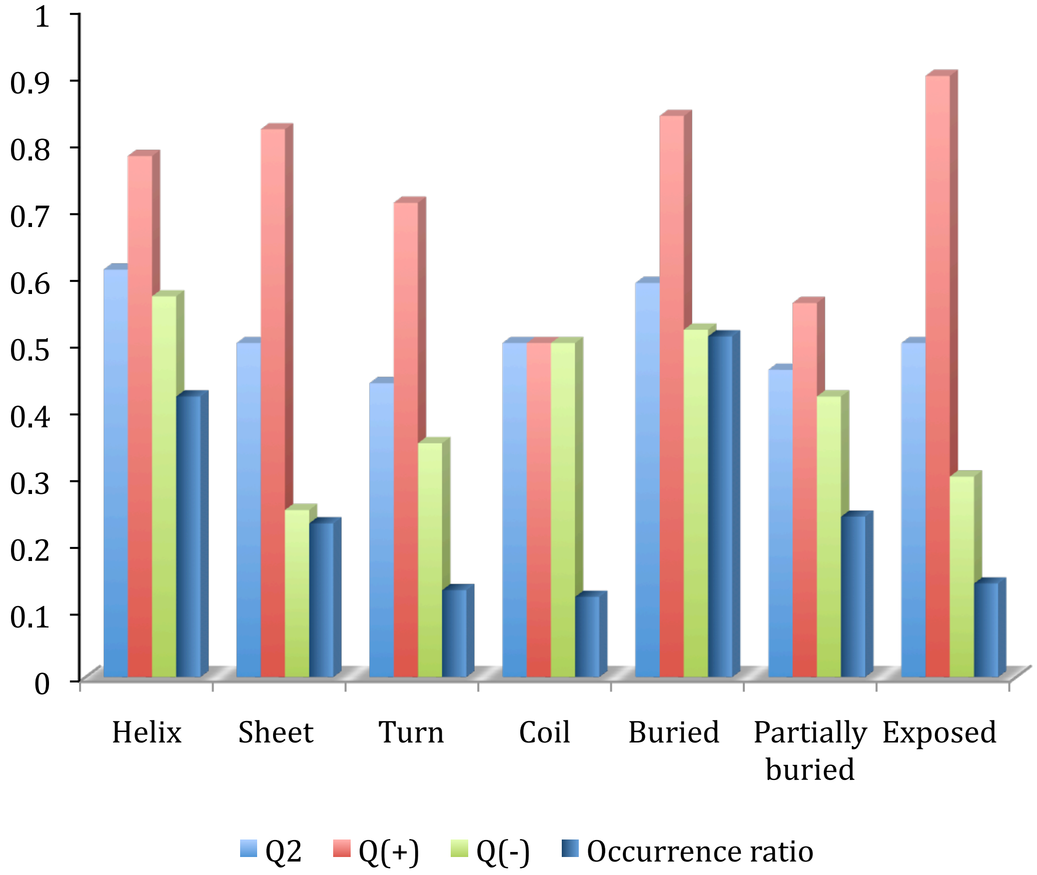


Figure 3.5 Classification accuracies of the optimum SVM model for the *ΔΔG* signs upon mutations of the single mutants in the test set according to the secondary structure allocation and the accessible surface area of the mutated residue. + and - : the indexes were evaluated for positive and negative *ΔΔG* signs. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class.

3.3. Discussion

3.3.1 Performance of the predictor according to the mutation type

The prediction accuracies of the *ΔΔG* real values varied for each mutation type, showing that the regression SVM model better learned the effects of some mutation types in comparison to others. The effects caused for charged/charged, charged/apolar and polar/charged mutations seem more complex. Therefore, the complete patters of such interactions are not contained in the mutant dataset or they cannot be successfully described from the sequence. In fact, sequence information partially describes the effect of salt-bridge and polar-polar interactions at the protein surface that should be better characterized using 3D structure details. Salt-bridge and hydrogen-bridge interactions at protein surface of charged and polar residues usually appear at long-ranges. Despite being separated by long stretches of polypeptide in the primary sequence, surface groups lie next to each other in space. In turn, at the protein core, interactions often occur among residues at short range. Consequently, these interactions are very difficult to model from a sequence framework. On the contrary, hydrophobic interactions at protein core mainly appear at short-range in the sequence. For mutations in the protein core, the residue size rather than polarity (apolar/apolar mutations), may cause an unfavorable packing energy due to the rigidity of surrounding residues or, alternatively, the substituting residues themselves may be forced into unfavorable rotational isomers. Some surroundings of the mutation positions may be readily deformable or compensate effects if occurs no net packing energy change [11]. Complex 3D environment-dependent interactions take place also in the protein core, which can be only accessible in some extend from a primary structure approximation.

In the case of helix mutations, the low prediction performance might be related to the fact that destabilization of helix structures is caused by variations of complex residue-residue and residue-solvent interaction patterns at the protein surface. The low correlation for mutations allocated at turn structures could be related to the variability and complex nature of turn regions in proteins and the low statistical significance of this group of mutations being 9% of the dataset. It is noteworthy that mutations of charged residues, mainly allocated at protein surface, were also predicted with low accuracy (Figure 3.2).

3.3.2 Prediction of protein real *ΔΔG* values upon mutations

Concerning the prediction of real values of change of Gibbs free energy change of proteins, several models used protein 3D structure information of large datasets (>1000) and no more than 60% of validation data variances were described [12-15]. AGADIR [16, 17] or FOLDEF was reported by Guerois et al. [17] for predicting conformational stability of more than 1000 mutants with crossvalidation accuracy about 60%. Zhou and Zhou method D-FIRE [14] is based on distance-scaled, finite ideal-gas reference state that improved structure-derived potentials of mean force for structure selection and stability prediction. Their model, with 3D protein structures from a database of 895 large-to-small mutations, described 0.45% of crossvalidation data variance. Borner and Abagyan [15] developed a model to predict both geometry and relative stability of point mutants that could be used for arbitrary mutations. An empirical energy function, which includes energy contributions of the folded and denatured proteins, and the prediction of a side chain mutant, was fitted to a training set consisting in a half of a diverse set of nearly 2000 experimental stability values for single point mutations. The prediction method was then tested on the remaining half of the experimental data, giving a covariance of 0.66 for 97% of the test set.

Furthermore, machine learning algorithms in combination with sequence or/and 3D information have been applied to solve the protein conformational stability problem [18-21]. Capriotti et al. [9, 10, 23] described the implementation of ANNs and SVMs predictors of *ΔΔG* upon mutations using sequences and 3D structures of more than 1000 mutants. As predictor inputs they used a combination of experimental condition data (pH and Temperature), specific mutated residue and sequence environment information. Their “best” sequence-based model explained a discrete

0.38% value of crossvalidation data variance. It is noteworthy that our optimum *AASA*-SVM overcomes the previous sequence-based models of Capriotti et al. [23]. It should be also noticed that they used a redundant dataset that over-estimate the predictor performance.

Recently, Huang et al. [18] reported the iPTREE-STAB server to discriminating the stability of proteins (stabilizing or destabilizing) and predicting their stability changes upon single amino acid substitutions from amino acid sequence. The predictor was trained with a dataset of 1859 non-redundant single point mutations of 64 proteins. The prediction of real $\Delta\Delta G$ values is mainly based on regression tree using three neighbouring residues of the mutant site along N- and C-terminals. Their method showed a crossvalidation correlation of 0.70 for predicting protein stability changes upon mutations, which is similar to our results. Other recent report by Cheng et al. [19] referred to the prediction of single mutant real $\Delta\Delta G$ values and signs by SVM predictors trained with information from three different encoding schemes: sequence, structure, and combined sequence and structure. In this case, the prediction of real $\Delta\Delta G$ values was higher than our results, having crossvalidation correlation coefficients of 0.75 and 0.76 for the sequence- and structure-based predictors, respectively.

3.3.4 Classification of protein conformational stability

Taking into account the classification models of protein stability change upon mutations using large and diverse mutant data, our classification model overcomes the optimum reported by Capriotti et al. [23] using sequence information. Despite they reported an overall accuracy about 77%, the correct predictions were drastically shifted towards unstable mutants with accuracy of 91% and for stable mutants the accuracy was lower about 46%. Such statistics reflect that their model nearly recognized all mutants as unstable, yielding overall adequate accuracy but inefficient discriminating ability. However, our classification model surpassed Capriotti's classifier by predicting unstable and stable mutants with accuracies over 70%. In this connection, our models yielded similar accuracy to our recent reports using SVM and 2D and 3D graph representations of protein sequences [24, 25], which identified both stable and unstable mutants with identical good accuracy over 70%. When Capriotti et al. [10] used 3D structure information, the highest overall classification accuracy was 80% but stable mutants were recognized with a low accuracy of 56%. In this context, our predictor, despite its primary sequence nature, is more adequate for the recognition of stable mutants.

Huang et al. [18] reported that iPTREE-STAB server was able to recognize unstable and stable mutants with overall crossvalidation accuracy about 82% and sensitivity and specificity were about 75.3% and 84.5%, respectively. The best sequence-based SVM classifier reported by Cheng et al. in Ref. 19 had an overall accuracy about 84% but it discriminated between unstable and stable mutants with accuracies about 90% and 71% respectively, in crossvalidation experiment. This result shows that the predictor is unbalanced and tends to recognize stable mutants with lower accuracy ($Q(+)$=0.71), which is equal to the reported value by our SVM classifier. It is noteworthy to mention that none of these predictors are able to handle multiple-point mutants.

In another recent report, Parthiban et al. [20] implemented a distance-dependant pair potential and torsion angle potential to compare predicted stabilizing energies with experimental values from thermal and chemical denaturation experiments. The derived force fields yielded a correlation of 0.77 and more than 80%

classification accuracy in crossvalidation for chemical denaturation. For thermal denaturation the force field yielded a correlation of 0.78 with a prediction efficiency of 84.65%.

3.3.5 Test set predictions

The low prediction result for the test set suggests that our sequence-based approach could be somehow limited due to the complexity of the stabilizing-destabilizing interactions in proteins. In addition, supervised learning of a predictor should have a training dataset with a complete description of the modeled phenomena. Training dataset should be complemented as more experimental conformational stability studies are published and collected in the Protherm database [8]. Probably more experimental measurements on protein conformational stability are needed to increase the outcome of machine learning approaches. Nevertheless, our method yields fast predictions of the stability of protein sequences. Beyond the agreeable results obtained in crossvalidation experiments and the modest results for the test set, the major advantage of our approach is the capability of our classifier to predict stability's changes upon double or multiple mutations. The predictor is online available at http://gibk21.bse.kyutech.ac.jp/llamosa/ddG-AASA/ddG_AASA.html.

3.3.6 Model's interpretation

Interestingly, relevant amino acid/residue properties appear weighting the optimum *AASA* vectors: three structural (N_m, f and ASA_N), one secondary structure-related (P_B), two physico-chemical (P and H_t) and three thermodynamic (G_{hN}, ΔG_C and ΔCp_h) properties. These relevant autocorrelations were found at lags from 7 to 15 medium to large range interactions on the sequence. The occurrence in our models of structural, secondary structure-related, physico-chemical and thermodynamic properties, reveals the complexity of the interactions ruling protein stability, which are better addressed by a multifactor approach.

Distributions of structural properties at lags ranging from 11 to 15 reflect the significance of an adequate amino acid frame at large ranges in the primary structure, resembling certain polypeptidic structural pattern. The number of medium range contacts (N_m) is a property that contains information of tridimensional proximities of residues in space. The property N_m also appeared weighting optimum autocorrelation vectors in a neural network implemented for modeling the conformational stability of chymotrypsin inhibitor 2 mutants [26]. Shape-related amino acid property, flexibility (*f*), appears relevant at autocorrelations of large range encoding the distribution of freedom degrees on the sequence. Re-accommodation of residue side-chains is a critical step in protein folding after amino acid substitution. Mutations may cause an unfavorable packing energy due to the rigidity of surrounding residues or, alternatively, the substituting residues themselves may be forced into unfavorable rotational isomers. Similarly, some surroundings of mutation positions may be deformable or some effects should be compensated to do not change packing energy [11]. The property *f* has appeared weighting optimum *AASA* vectors for modeling conformational stability and functional variations upon mutations of gene V protein [22] and ghrelin receptor [27]. Another main structural property is the solvent accessible surface area for native protein (ASA_N), which is a measure of the number of amino acid atoms interacting with solvent molecules in the native state. Interestingly, solvent-accessible surface area was reported by Gromiha et al. [28] among the most linearly correlated properties with $\Delta\Delta G$ for a diverse set of protein mutants. In this connection, one of the simplest and most widely used models for calculating hydration heat capacity in proteins is the solvent-accessible surface area model [29].

The high impact of β-structure tendency strongly suggests that optimum secondary structure pattern is another key factor for a stable tertiary conformation. Point mutations studies have highlighted the role of secondary structure propensities in protein stability. By manipulating favorable and unfavorable secondary structure propensities at certain positions in a protein can produce significant variations in stability [12]. In fact, we recently reported that secondary structure propensities are important in a neural network model of the conformational stability of gene V protein mutants [22].

Hydrophilicity/hydrophobicity related properties such as polarity (P) and thermodynamic transfer hydrophobicity (H_t) are important for predicting protein conformational stability according to our optimum SVM models. The important autocorrelations of such properties appears at medium lags. Hydrophilic interactions between amino acid residues at protein surface usually appear at medium and long ranges. Despite being separated by long stretches of polypeptide in the primary sequence, surface groups lie next to each other in space. On the contrary, hydrophobic interactions at protein core mainly appear at shorter range in the sequence. The P and H_t properties were previously found relevant for neural network modeling of conformational stability of human lysozyme mutants [21]. The autocorrelation vectors weighed by these properties encoded the role of hydrophylic interactions on the surface and hydrophobic interactions on the core to maintain protein folding and stability. Furthermore, hydrophobic patches frequently appear on the protein surface, defined as clusters of neighboring apolar atoms accessible on a given protein surface [30]. The hydrophobic part of the solvent-accessible surface of a typical monomeric globular protein consists of a single, large interconnected region formed from faces of apolar atoms and constituting approximately 60% of the solvent-accessible surface area [31]. At the light of these facts, the combination of hydrophilicity/hydrophobicity and solvent-accessible surface properties could encode hydrophobic patches patterns of protein mutants.

Thermodynamical properties that measure unfolding and hydration processes of proteins (unfolding Gibbs Free energy change of side-chain (ΔG_C), Gibbs Free energy change of hydration for native protein (G_{hN}) and hydration heat capacity change (ΔCp_h)) are relevant to model protein conformational stability. In a previous report, the property G_{hN} was relevant in SVM models of the functional variations upon mutations of ghrelin receptor [27], meanwhile ΔG_C and ΔCp_h properties were important in neural network modeling of human lysozymes conformational stability [21]. ΔCp_h measurements in proteins mean the variation of heat capacity (Cp), which is consequence of the hydration of amino acid groups. Considering that protein unfolding usually has a positive ΔCp, polar groups hydration is accompanied by a decrease in Cp, meanwhile apolar groups hydration increases this magnitude [32]. In this sense, Makhatadze and Privalov [33] found a good relation between ΔCp_h and surface area. In turn, G_{hN}, is a measure of spontaneity of the hydration process. Free energy has a direct relationship to a primary observable, the equilibrium constant K, through $\Delta G = -kTlnK$, which describes the balance between enthalpy and entropy. Makhatadze and Privalov [33] showed that the compact native state of a protein is stabilized by the enthalpic interactions between internal groups. Hydration effects are clearly significant for protein unfolding, evidence showed that hydration is the major effect [33]. The strongest current evidence is that it can be accounted for heat capacity change of unfolding for many proteins by adding up hydration contributions from individual residues [33]. In resume all those thermodynamical properties are related with unfolding denaturation mechanism hypothesis. For denaturation process of globular proteins, Privalov and Gill [34] pointed out the hydration equilibrium, polar interactions between solvent and polar residues in the protein, as the main cause of unfolding while hydrophobic interactions in the protein core contribute to keep the folded state.

3.4. Conclusions

GA-SVMs yielded a good classification model for the conformational stability of protein mutants describing nearly 80% of correct classifications in crossvalidation experiment. The regression model described nearly 50% of crossvalidation data variance. Despite low test set prediction accuracy, stable single and double point mutants were recognized with adequate accuracies about 70%. Optimum *AASA* vectors, selected by GA-SVM approach, showed that conformational stability model depends on a combination of structural, secondary structure-related, physico-chemical and thermodynamical properties mainly associated with protein hydration process.

3.5. References

1. Bauknecht H, Zell A, Bayer H, Levi P, Wagener M, Sadowski J, Gasteiger J: **Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists**. *J Chem Inf Comput Sci* 1996, **36**:1205–1213.

2. Moran PAP: **Notes on Continuous Stochastic Processes**. *Biometrika* 1950, **37**: 17–23.

3. Geary RF: **The contiguity ratio and statistical mapping**. *The Incorporated Statistician* 1954, 5:115–145.

4. Moreau G, Broto P: **Autocorrelation of a topological structure: A new molecular descriptor**. *Nouv J Chim* 1980, **4**:359–360.

5. Moreau G, Broto P: **Autocorrelation of Molecular Structures: Application to SAR Studies**. *Nouv J Chim* 1980, **4**:757–764.

6. Wagener M, Sadowski J, Gasteiger J: **Autocorrelation of Molecular Properties for Modelling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks**. *J Am Chem Soc* 1995, **117**:7769–7775.

7. (a) K. Nakai, A. Kidera, M. Kanehisa: **Cluster analysis of amino acid indices for prediction of protein structure and function**. *Protein Eng* 1988, **2**:93–100. (b) K. Tomii, M. Kanehisa: **Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins**. *Protein Eng* 1996, 9:27–36. (c) S. Kawashima, M. Kanehisa: **AAindex: amino acid index database**. *Nucleic Acids Res* 2000, 28:374.

8. Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34**:D204– D206.

9. Capriotti E, Fariselli P, Calabrese R, Casadio R: **Prediction of protein stability changes from sequences using support vector machines**. *Bioinformatics* 2005, **21**:54 –58.

10. Capriotti E, Fariselli P, Calabrese R, Casadio R: **Prediction of protein stability changes from sequences using support vector machines**. *Bioinformatics* 2005, **21**:54 –58.

11. a) Sandberg WS, Terwilliger TC: **Energetics of repacking a protein interior**. *Proc Natl Acad Sci USA* 1991, 88:1706–1710. b) Sandberg WS, Terwilliger TC: **Engineering multiple properties of a protein by combinatorial mutagenesis**. Proc. Natl Acad Sci USA 1993, 90:8367–8371.

12. Guerois R, Nielsen JE, Serrano L: **Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations**. *J Mol Biol* 2002, **320**:369 –387.

13. Zhou H, Zhou Y: **Stability Scale and Atomic Solvation Parameters Extracted From 1023 Mutation Experiment**. *Proteins* 2002, **49**:483–492.

14. Zhou H, Zhou Y: **Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction**. *Protein Sci* 2002, **11**:2714–2726.

15. Bordner AJ, Abagyan RA: **Large-Scale Prediction of Protein Geometry and Stability Changes for Arbitrary Single Point Mutations**. *Proteins* 2004, **57**:400–413.

16. Lacroix E, Viguera AR, Serrano L: **Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters**. *J Mol Biol* 1998, **284**:173–191.

17. Munoz V, Serrano L: **Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm–Bragg and Lifson–Roig formalisms**. *Biopolymers* 1997, **41**:495–509.

18. Huang L-T, Gromiha MM, Ho S-Y: **iPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations**. *Bioinformatics* 2007, 23-1292–1293.

19. Cheng J, Randall A, Baldi P: **Prediction of protein stability changes for single-site mutations using support vector machines**. *Proteins* 2006, 62:1125–1132.

20. Parthiban V, Gromiha MM, Hoppe C, Schomburg D: **Structural analysis and prediction of protein mutant stability using distance and torsion potentials: Role of secondary structure and solvent accessibility**. *Proteins* 2006, 66:41–52.

21. Caballero J, Fernández L, Abreu JI, Fernández M: **Amino Acid Sequence Autocorrelation Vectors and Ensembles of Bayesian-Regularized Genetic Neural Networks for Prediction of Conformational Stability of Human Lysozyme Mutants**. *J Chem Inf Model* 2006, **46**:1255–1268.

22. Fernández L, Caballero J, Abreu JI, Fernández M: **Amino Acid Sequence Autocorrelation Vectors And Bayesian-Regularized Genetic Neural Networks For Modeling Protein Conformational Stability: Gene V Protein Mutants**. *Proteins* 2007, **67**:834–853.

23. Capriotti E, Fariselli P, Casadio R: **A neural-network-based method for predicting protein stability changes upon single mutations**. *Bioinformatics* 2004, **20**:63–68.

24. Fernández M, Caballero J, Fernández L, Abreu JI, Acosta G: **Classification of conformational stability of protein mutants from 2D graph representation of protein sequences using support vector machines**. *Mol Simulat* 2007, **33**:889–896.

25. Fernández M, Caballero J, Fernández L, Abreu JI, Acosta G: **Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines**. *Proteins* 2008, **70**:167–175.

26. Fernández M, Abreu JI, Caballero J, Garriga M, Fernández L: **Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks**. *Mol Simulat* 2007, **13**:1045–1056.

27. Caballero J, Fernández L, Garriga M, Abreu JI, Collina S, Fernández M: **Proteometric Study of Ghrelin Receptor Function Variations upon Mutations using Amino Acid Sequence Autocorrelation Vectors and Genetic Algorithm-Based Least Square Support Vector Machines**. *J Mol Graph Model* 2007, **26**:166–178.

28. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A: **Relationship Between Amino Acid Properties and Protein Stability: Buried Mutations.** *J Prot Chem* 1999**, 18**: 565–578.

29. Prabhu NV, Sharp KA: **Heat capacity of proteins**. *Annu Rev Phys Chem* 2005, **56**: 521–548.

30. Lijnzaad P, Argos P: **Hydrophobic patches on protein subunit interfaces: characteristics and prediction**. *Proteins* 1997, **28**:333–343.

31. Eisenhaber F, Argos P: **Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation**. *Protein Eng* 1996, **9**:1121–1133.

32. a) Makhatadze GI, Privalov PL: **Heat capacity of proteins. I. Partial molar heat capacity of individual amino acid residues in aqueous solutions: hydration effect**. *J Mol Biol* 1990, **213**:375–384 b) Makhatadze GI, Privalov PL: **Partial molar heat capacity of the unfolded polypeptide chain of proteins: protein unfolding effects**. *J Mol Bio* 1990, **213**:385–391.

33. Makhatadze GI, Privalov PL: **Hydration effects in protein unfolding**. *Biophys Chem* 1994, **51**:291–309.

34. Privalov PL, Gill SJ: **Stability of Protein Structure and Hydrophobic Interaction.** *Adv Prot Chem* 1988, **39**:191–234.

CHAPTER 4. MODELING OF KINASE INHIBITION

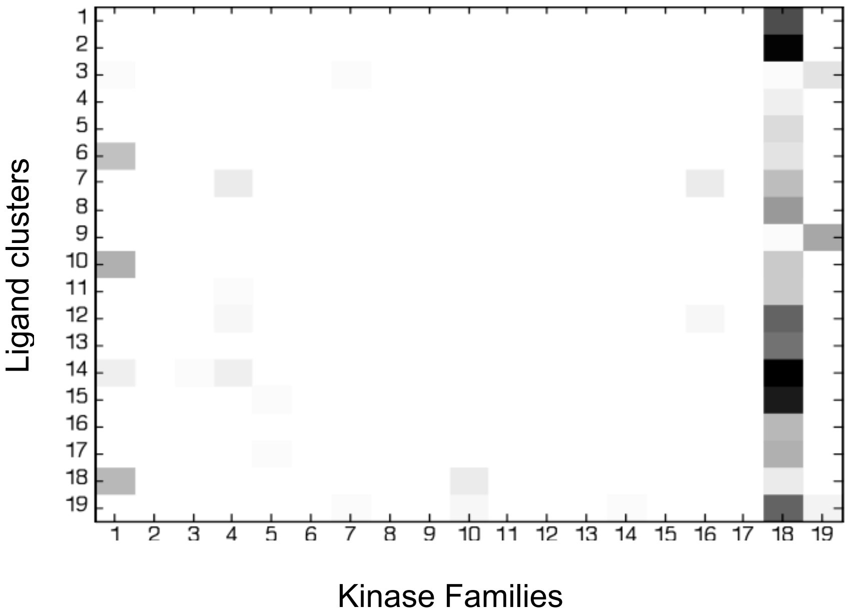
4.1. Introduction

This chapter presents the modeling of the ligand dataset with reported inhibitory activities towards 62 kinases collected from ProLiNT database [1]. Structural fragments and topological autocorrelations features were computed from protein sequences and 2D graphs of ligands. The dataset was divided into training (80% dataset) and test sets (20% dataset) by k-means clustering. Five clusters were generated and cases were homogeneously added to training and test sets by selecting instances from each cluster according to cluster's sizes. Subsequently, Support Vector Machines (SVMs) model was used to classify hypothetical complexes into stable and unstable classes. Predictor's optimization was carried out by three-fold-out (TFO) crossvalidation. The training set was divided into three subsets: two subsets were used for training the classifier and the rest subset was then predicted. This process was repeated until all the subsets were predicted. The ability of the predictor to recognize new ligands and targets was evaluated in the test set. The performance for different kinase families and inhibitor scaffolds were also evaluated.

4.2. Results

4.2.1 Clustering analysis

Selectivity of the 19 kinase families was firstly explored by cluster analysis. Two sets of variables, structural fragments and 2D autocorrelation vectors were calculated for the 1200 active ligands on the dataset. Afterwards, k-means clustering algorithm yielded 19 ligand clusters (same as the number of kinase families) for each descriptor type. In order to evaluate the efficiency of kinase selectivity in the two descriptor's sets, we matched ligand clusters with the different kinase families.



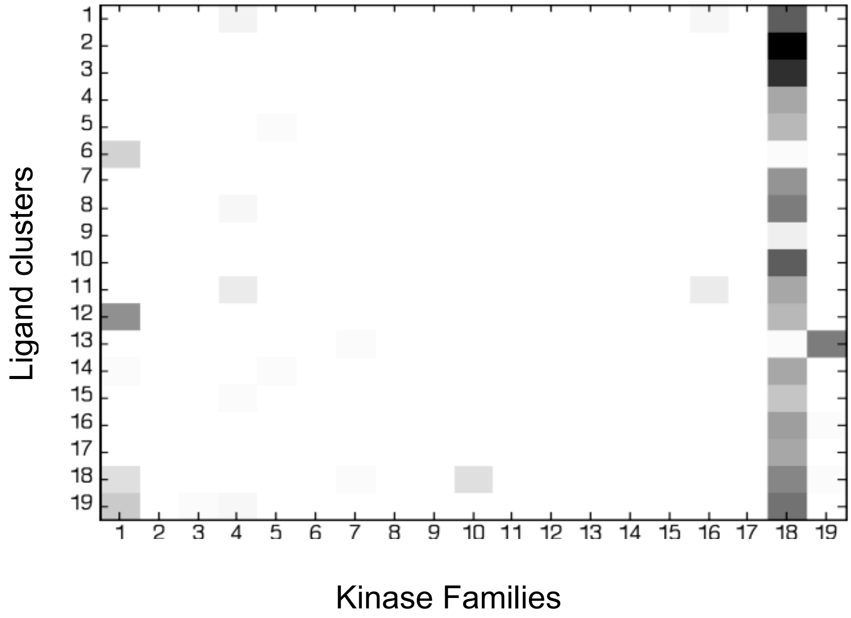


Figure 4.1. Density map representing occurrence ratios of kinase families on active ligand clusters obtained by atom-center fragment descriptors (A) and 2D autocorrelation vectors (B). Kinase families: (1) AGC Ser/Thr protein kinase, (2) CAMK Ser/Thr protein kinase, (3) CAMP-dependent kinase regulatory chain, (4) CMGC Ser/Thr protein kinase, (5) cyclin, (6) DCK/DGK, (7) herpes virus thymidine kinase, (8) PDGF/VEGF growth factor, (9) PI3/PI4-kinase, (10) PI3K p85 subunit, (11) PP1 inhibitor, (12) phosphoglycerate kinase, (13) phosphorylase b kinase regulatory chain, (14) STE Ser/Thr protein kinase, (15) Ser/Thr protein kinase, (16) TKL Ser/Thr protein kinase, (17) thymidine kinase, (18) Tyr protein kinase, and (19) atypical kinase.

Distributions of the two sets of 19 ligand clusters on the 19 kinase families are shown in Figure 4.1. The cluster distributions depict different patterns. Tyrosine kinase family was the most populated in both graphs but distributions of the ligand clusters on this kinase family differ. In addition, we built two dendrograms by clustering kinase families according to the distributions of the ligand clusters in the families. Figure 4.2 depicts dendrograms for both descriptor types showing 5 clusters. The dendrogram in Figure 4.2A exhibits similar distributions of kinase families in the clusters in comparison with the dendrogram in Figure 4.2B. The five clusters in Figure 2A represent 4 kinase families and a group of 15 kinase families. The single-family clusters are: Tyrosine kinase, Cyclin, AGC Ser/Thr protein kinase and atypical kinases. In turn, Figure 4.2B depicts three single-family clusters: Tyrosine kinases, AGC Ser/Thr protein kinases and atypical kinases; one cluster with two families: CMGC Ser/Thr protein kinase and TKL Ser/Thr protein kinase and a several ten-families cluster. The distributions of the kinase families in the several ten-families clusters differ at lower squared Euclidean distances.

The clustering points out that independently allocated kinase families are inhibited by different chemical scaffolds. At the same time, families in multiple clusters could share some active ligand similarity depending on each encoding scheme. Especially, the allocation of CMGC Ser/Thr protein kinase and TKL Ser/Thr protein kinase in the same cluster in Figure 4.2B suggests a similar inhibition scaffold for these families according to the topological approach.

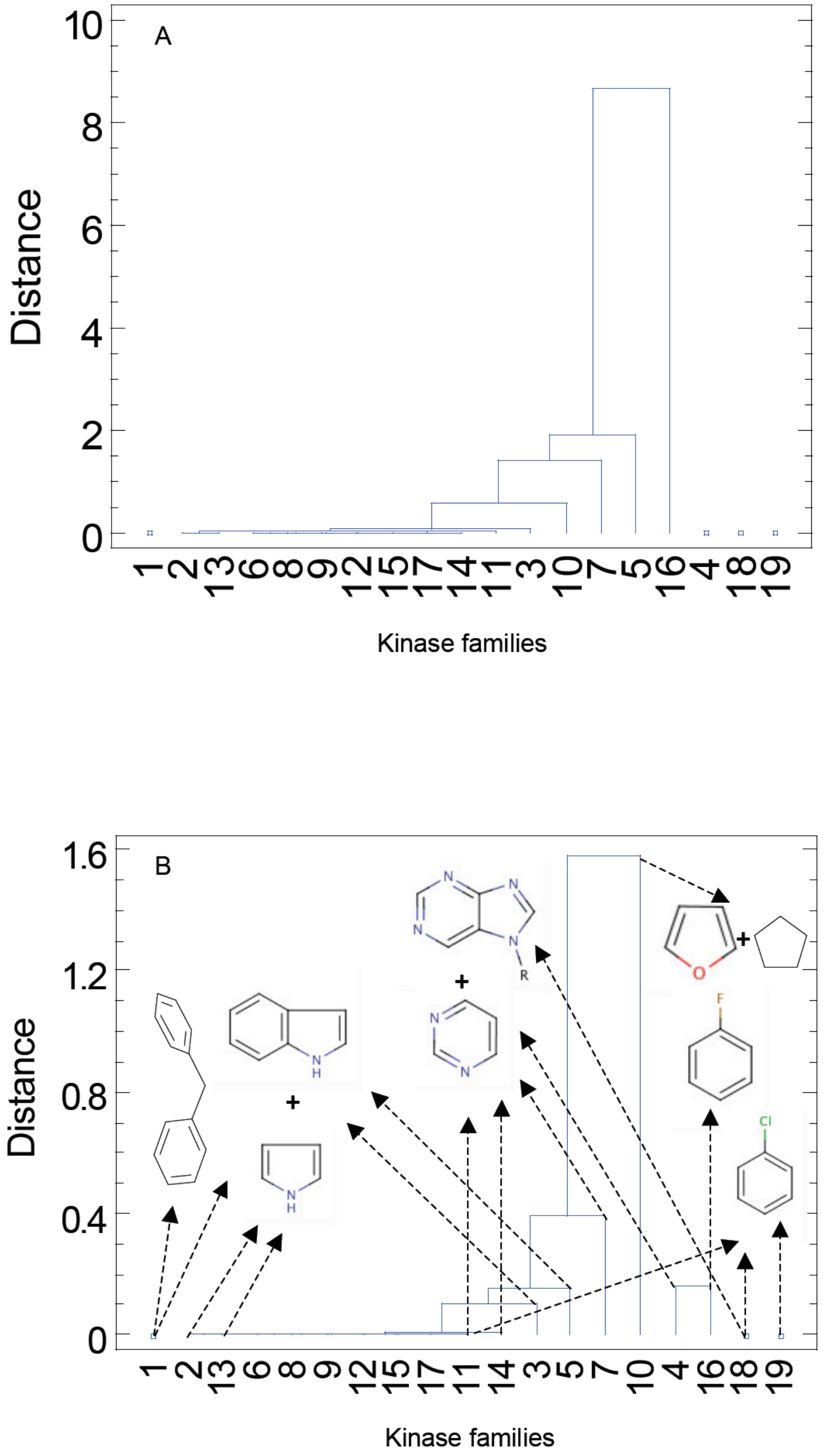


Figure 4.2. Dendrograms of 5 clusters of kinase families according to occurrences ratios on active ligand clusters obtained by fragment descriptors (A) and 2D autocorrelation vectors (B). Kinase families: (1) AGC Ser/Thr protein kinase, (2) CAMK Ser/Thr protein kinase, (3) CAMP-dependent kinase regulatory chain, (4) CMGC Ser/Thr protein kinase, (5) cyclin, (6) DCK/DGK, (7) herpes virus thymidine kinase, (8) PDGF/VEGF growth factor, (9) PI3/PI4-kinase, (10) PI3K p85 subunit, (11) PP1 inhibitor, (12) phosphoglycerate kinase, (13) phosphorylase b kinase regulatory chain, (14) STE Ser/Thr protein kinase, (15) Ser/Thr protein kinase, (16) TKL Ser/Thr protein kinase, (17) thymidine kinase, (18) Tyr protein kinase, (19) atypical kinase.

We also evaluated the ability of a fragment and its topological descriptors to distinguish between active and inactive chemical scaffolds in the dataset. The mean squared differences between the normalized descriptors of active and inactive inhibitors were 0.361 and 0.334 from the fragment- and topological-based approaches respectively. These values significantly differed from the mean squared differences of 100 scrambled partitions of active and inactive ligands, which were about 10^{-2} for both descriptor types. Topological autocorrelation descriptors differentiate slightly better between active and inactive inhibitors in comparison with the fragment approach.

4.2.2 Structural Fragments classifier

SF and *SSF* descriptors computed for the kinase sequences and the 2D structural sketches of the inhibitors were combined into a single feature matrix by simple concatenation of target and ligand descriptors blocks. The dataset was separated into a training set with 2696 inhibition complexes (80%) and a test set with 899 inhibition complexes (20%).

In a first attempt, we implemented a linear kernel but the highest crossvalidation accuracy was only about 65%. Then, adjusting SVM parameters throughout a TFO crossvalidation in a grid search yielded a nonlinear classifier with crossvalidation results shown in Table 4.1. An overall TFO crossvalidation accuracy of 78% for the classification of inhibition complexes was achieved with a correlation coefficient *Cr*=0.53. It is noteworthy that crossvalidation accuracies for recognizing stable *Q(+)*=0.75 and unstable inhibition complexes *Q(-)*=0.78 are equivalent to the overall accuracy achieved. Taking into account that the predictor was trained with structural fragment information from targets and ligands, these accuracies about 80% for recognizing stable and unstable inhibition complexes are adequate.

Table 4.1. Crossvalidation and prediction statistics for the training and test sets according to the fragment SVM model for the classification of the stability of kinase inhibition complexes.

Experiment	*Q2*	*Q(+)*	*Q(-)*	*P(+)*	*P(-)*	*Cr*
Training set crossvalidation	0.78	0.75	0.79	0.65	0.86	0.53
Test set prediction	0.78	0.77	0.77	0.65	0.88	0.55

+ and - : the indexes were evaluated for “stable” (IC_{50} < 1 μM) and “unstable” (IC_{50} > 1 μM) kinase inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *P(s)* is the number of correct prediction for class s/all prediction made for s; *Q(s)* is the number of correct prediction for class s/observed in class s; *Cr* is Matthews’s correlation coefficient. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction and *P(+)* and *P(-)* are precision scores.

4.2.3 Topological autocorrelation classifier

Topological features were computed from the primary sequence of the kinases and the 2D sketch of the inhibitor structure descriptors. Similar to the fragment-based classifier, the dataset was separated into training and test sets in the same way.

A linear kernel could only produce a poor performance (67% accuracy) in a cross-validated training. Then, adjusting SVM parameters throughout a TFO crossvalidation in a grid search yielded an optimum nonlinear SVM classifier with crossvalidation results shown in Table 4.2. In Table 4.2 complex-wise statistics refers to the basic crossvalidation experiments in which only nonredundant complexes were included. In this case, an overall complex-wise TFO crossvalidation accuracy of 82% was achieved for the classification of inhibition complexes with a correlation coefficient *Cr*=0.63. Crossvalidation accuracies to identify stable *Q(+)*=0.85 and unstable inhibition complexes *Q(-)*=0.81 resulted similar to the overall accuracy. These accuracies and the correlation coefficient are higher than the statistics reported in Table 4.1 for the fragment-based predictor.

Table 4.2. Crossvalidation and prediction statistics for the training and test sets according to the optimum topological SVM model with for the classification of the stability of kinase inhibition complexes. SVM parameters were σ^2=0.091 and C=1.36.

Experiment		*Q2*	*Q(+)*	*Q(-)*	*P(+)*	*P(-)*	*Cr*
Training set crossvalidation	Complex-wise	0.82	0.85	0.81	0.69	0.92	0.63
	Ligand-wise	0.82	0.84	0.81	0.68	0.91	0.62
	Kinase-wise	0.65	0.75	0.60	0.48	0.83	0.32
Test set prediction	Complex-wise	0.81	0.87	0.78	0.67	0.92	0.62
	Ligand-wise	0.82	0.91	0.77	0.69	0.94	0.66

+ and - : the indexes were evaluated for "stable" (IC_{50} < 1 μM) and "unstable" (IC_{50} > 1 μM) kinase inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *P(s)* is the number of correct prediction for class s/all prediction made for s; *Q(s)* is the number of correct prediction for class s/observed in class s; *Cr* is Matthews's correlation coefficient. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction and *P(+)* and *P(-)* are precision scores.

Fragment descriptors are more intuitive and easy to interpret, but they only account for substructure occurrences on the structure and lack the information of connectivity and sequence order. In turn, 2D autocorrelation vectors account for property distributions on the topological structure accounting for atom arrangements in the bi-dimensional molecular sketch and amino-acid residue distributions along the protein sequence. In view of the training set results, we conclude that autocorrelation approach outperforms the fragment-based classifier and it is more convenient for modeling kinase inhibition.

In addition, the stability of the optimum topological model to recognize new ligands and kinases was evaluated. We performed two additional crossvalidation experiments in which kinase complexes sharing similar ligands were kept in the same data subset during crossvalidation, we called this experiments ligand-wise crossvalidation. Similarly, we run another crossvalidation in which complexes of the same kinase were kept in the same subset and this was called kinase-wise crossvalidation. The results of these experiments are reported in Table 4.2, overall accuracy for ligand-wise crossvalidation was 82%, illustrating that the model correctly classifies the affinity of novel ligands towards existing kinases. All further reported statistical analysis for kinase families (Figures 4.3 and 4.4) and ligand chemotypes (Table 4.1 Appendix) were performed on ligand-wise crossvalidation. Interestingly, the kinase-wise 10-fold-out crossvalidation in Table 4.2 showed that the model differentiated complexes of new kinase with overall accuracy about 65% and stable complexes with accuracy of 75%. This result, although discrete, is noteworthy taking into account that when removing highly represented kinases from the training subset also large series of inhibitors are left out. This fact corroborates the relevance of the topological feature space to model kinase inhibition as well as the self-consistency of the optimum SVM model.

4.2.4 Performance of the optimum topological autocorrelation classifier for different kinase families and chemotypes.

Kinase inhibition dataset includes inhibitory activities of a diverse chemical space towards 19 kinase families. It is very interesting to analyze the optimum classifier performance for each kinase family in the dataset. The classification accuracies of SVM predictor for each kinase family are shown in Table 4.3. The predictor performance was very homogeneous to all families. The overall accuracies for the recognition of stable and unstable inhibition complexes were higher than 67% for all but one the kinase families. However, the classifier was unable to recognize stable inhibition complexes of seven protein kinase families with low occurrences of stable complexes in the crossvalidation experiments. This fact suggests that the training set information is highly diversified and generalization from one family to the others is difficult inside the training set. In addition to the low statistical significances of the stable inhibitors of these families in the dataset, another factor accounting for these low accuracies in crossvalidation experiments could be the complexity of the target-ligand interactions for these kinase families. In this regard, a recent review on QSAR modeling of binding affinities stated that conformational changes and binding site flexibility lead to the conclusion that similar analogs bind to the same binding site in different modes. Furthermore, the binding site residues in the ligand-protein interactions are not the same due to the difference in the flexible binding site residues [2].

Although our approach is alignment- and conformation-independent, identical or closely related inhibitor structures, which interact in different ways, could cause model failure for such inhibition complexes. However, one of the advantages of the alignment- and structure-free protein activity/function prediction methods is that they are less prone to be affected by protein folding or ligand's binding orientation. At the same time, for QSAR studies, when the binding mechanism and orientation are unknown and if a broad variety of targets are processed, it is usually accepted that more robust and accurate models can be derived from 2D-structure encoding frameworks rather than 3D detailed description of the molecules.

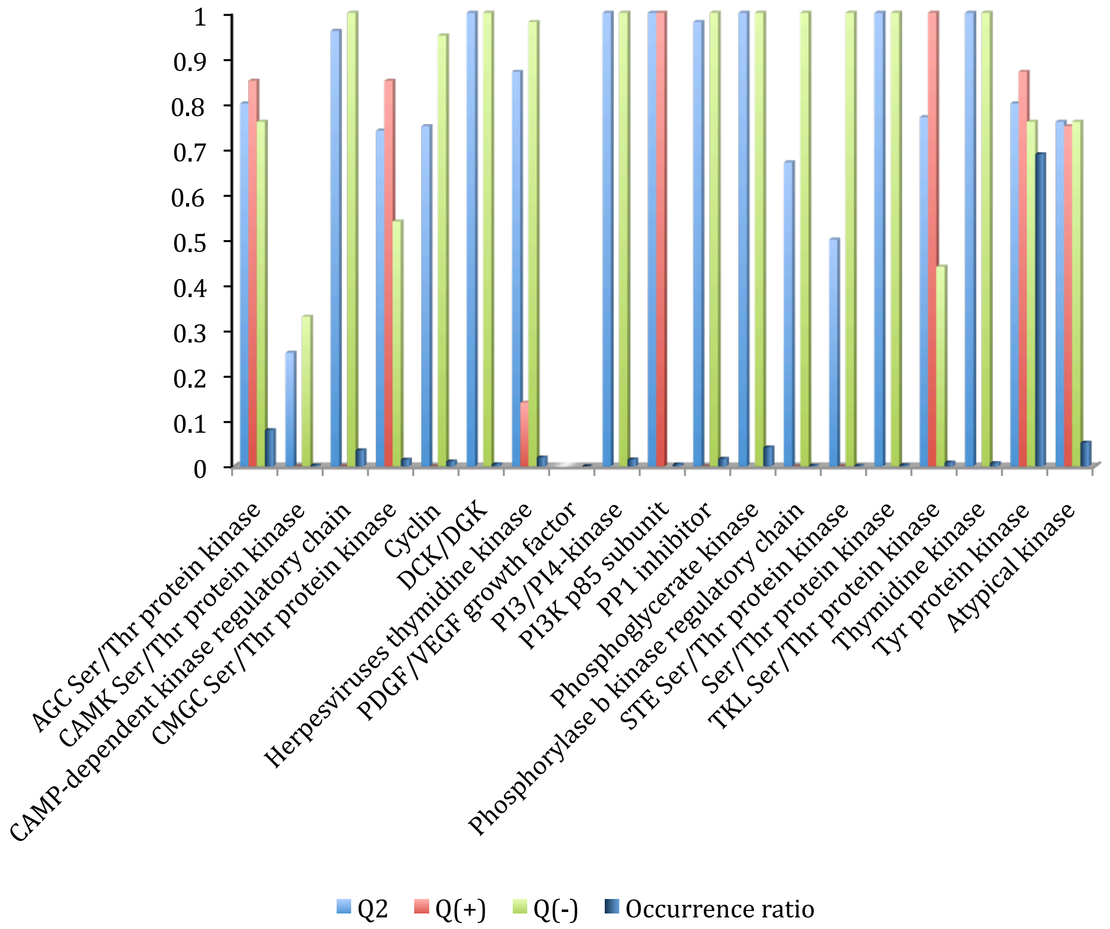


Figure 4.3. Ligand-wise TFO crossvalidation accuracies for the 19 kinase families in the training set according to the optimum topological SVM model. + and - : the indexes were evaluated for "stable" (IC_{50} < 1 µM) and "unstable" (IC_{50} > 1 µM) kinase inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class *s*/observed in class *s*. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction.

According to Figure 4.3, kinase families with the highest occurrence ratios on the training set exhibit crossvalidation accuracies higher than 80%. These kinase families are Tyr protein kinase, AGC Ser/Thr protein kinase, Phosphoglycerate kinase and CAMP-dependent kinase regulatory chain families with occurrence ratios of 68.81%, 7.97%, 4.15% and 3.52%, respectively. Among these families, tyrosine protein kinase has been most studied for targeting cancer. Tyrosine protein kinase directly participates in cell growth through the signal passing pathways. Five types of proteins participate in the growth control of mammalian cells: growth factors, growth factor receptors, intracellular transducers, nuclear transcription factors, and cell cycle control proteins. Some cell surface receptors have an extracellular ligand-binding domain attached to an integral protein tyrosine kinase in their cytoplasmic domain. These receptors transmit the growth signal by phosphorylating their tyrosine residues as well as one or more target of proteins, thus initiating a cascade of events [3]. It is widely accepted that protein tyrosine kinases play a fundamental role in cancer. That is why they became attractive therapeutic targets and it has provided impetus for an extensive effort to develop specific inhibitors of these enzymes as chemotherapeutic agents. An overall high performance (~80%) of our classifier in predicting the stable versus unstable complexes in this kinase family, therefore, has very useful practical implications to the inhibitor design problem.

We also analyzed the predictor behavior for different chemical subspaces on the inhibitor dataset. In this sense, 30 substructural templates were considered for comparing the classifier accuracy regarding the different chemotypes on the modeled chemical space (Table 4.1 Appendix). All the analyzed substructures showed overall accuracies about or higher than 70% and accuracies for separate classes were lower than 50% only for low affinity ligands bearing 1,3-dichlorobenzene substructure and high affinity ligands bearing m-methyltoluene. From this result we conclude that the classifier performed well over the chemical space represented by the kinase inhibitors in the training set.

4.2.5 Prediction of the test set

Crossvalidation accuracy gives an estimate of the internal consistency of the predictive models but a more realistic measurement of the prediction power can be achieved by predicting a blind test set. In Table 4.2, we show that the results from the topological-based classifier on such datasets also perform well and lead to an overall accuracy of about 81% for the complex-wise evaluation. This is promising considering the fact that the test set prediction accuracies were in the same range as obtained in crossvalidation experiments of the training set in Table 4.2, thus excluding the possibility of overfitting. Classification of test set also performed well for the topological models taking into account that test set accuracy of the fragment-based predictor was 78% (Table 4.1).

In addition, the ability of the topological predictor to recognize totally new ligands was estimated by evaluating the model accuracy for 462 out of the 899 kinase inhibition complexes in the test set for which ligands information was not available in the training set. This predictor correctly classified 82% of the 462 kinase complexes with totally new ligands. Remarkably, stable complexes were recognized with accuracy of 91% whilst the accuracy for the unstable complexes was 77%. This result showed that the optimum classifier not only correctly learned the kinase inhibition pattern, but that the learned pattern was adequately generalized to the test set, including totally new high-affinity ligands.

Figure 4.4 depicts classification results for each kinase family on the test set. All kinase families were classified with overall accuracies > 55%. Furthermore, classifier performance has accuracies over 80% for 16 out of the 18 kinase families in the test set. The model failed to recognize only high affinity ligands of CAMP-dependent kinase regulatory chain family. Thus, the information from the training set was successfully generalized and the prediction results were homogenous to the majority of the kinase families in the test set.

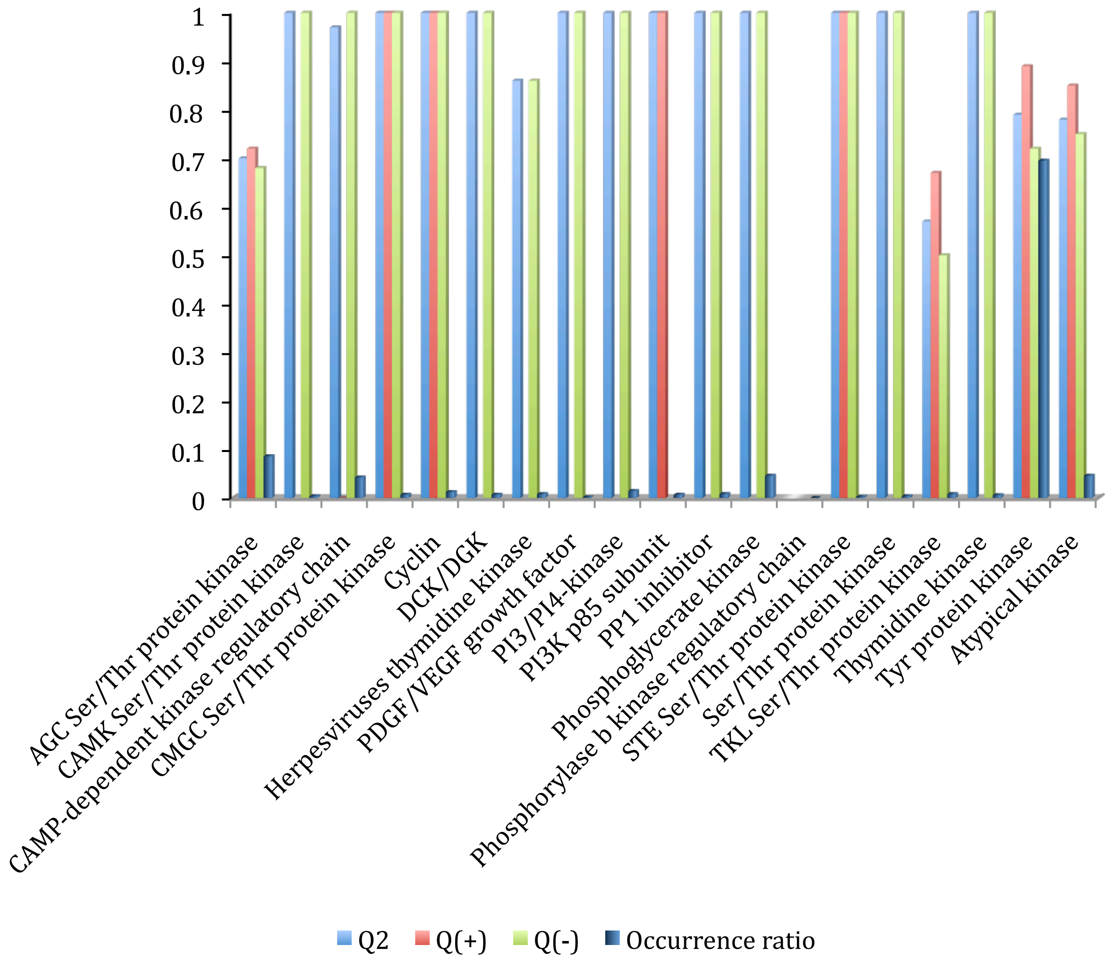


Figure 4.4. Ligand-wise prediction accuracies for the 19 kinase families in the kinase inhibitor in the test set according to the optimum SVM model. + and - : the indexes were evaluated for "stable" (IC_{50} < 1 μM) and "unstable" (IC_{50} > 1 μM) kinase inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction.

Similarly, the analysis of the classifier performance on the test set according to different chemotypes shows that the classifier attained similar performance to the training sets used in crossvalidation in Figure 4.4. All the inhibitor types have overall accuracies about or higher 50% and only 1,3-difluorobenzene has accuracy < 70%. Low affinity ligands bearing 1,3-diclorobenzene chemotypes were classified with low accuracies and the predictor failed to recognized low affinity 1,3-difluorobenzene derivatives. Despite of the low prediction performance for these chemotypes, the overall performance of the predictor on this blind set is adequate. The classifier recognized the inhibition pattern from different kinase families and also properly discriminated between stable and unstable inhibitor complexes belonging to several chemical subspaces in the test set.

The differential relevance of the topological autocorrelation space for modeling kinase inhibition was evaluated by sensitivity analysis [4]. The impact of each variable in the optimum model was estimated by measuring test set prediction accuracies for different modified feature matrices in which each autocorrelation vector at a time was replaced by a constant vector of same length. The magnitude of the importance of each input variable in the predictor was taken as the underperformance score defined as the ratio between *F-scores* (Eq. 2.11) for original and modified feature matrices. The sensibility analysis yields top-9 relevant autocorrelation vectors in descending order as follows: $AASA3H_t > AASA2R_a > AASA4R_a > AASA5R_a > AASA4ASA_N > AASA5ASA_N > ATS6v > ATS3e > ATS4e$. It is noteworthy that the most relevant inputs are kinase's autocorrelations of hydrophobicity/polarity-related properties such as thermodynamic transfer hydrophobicity (H_t), solvent-accessible reduction ratio (R_a) and solvent-accessibility area for native state (ASA_N), in combination with ligand's autocorrelations of atomic volumes and electronegativities on the 2D structure sketch.

4.3. Discussion

Inhibition of protein kinases can be broadly classified into three categories: ATP-competitive inhibition, substrate-competitive inhibition, and allosteric inhibition. Successful treatments of chronic myeloid leukemia and gastrointestinal stromal tumor with Gleevec [5] have recently drawn much attention because of its excellent selectivity and its ability to bind to a precise inactive conformation of Abl kinase. However, the emergence of drug-resistant mutants [6] and structural studies suggest that mutations in the kinase domain cause resistance to the Abl kinase inhibitor [7,8]. Other studies have also shown that some inhibitors can recognize specific inactive conformation of B-Raf (1UWH) and p38 (1W83), whereas some others can inhibit the active form of Abl kinase [9-11]. However, all of them have been shown to decrease or completely lose inhibitory activity towards some mutated kinase. In this sense, Thaimatta et al. [12] in the review of kinase inhibitors stated that the modulation of kinase activity has not been sufficiently exploited for therapeutic purposes. These authors suggested that inhibition of a single kinase may be insufficient to achieve a therapeutic benefit, and that promiscuous small-molecule kinase inhibitors or cocktails of inhibitors may be more promising than selective agents by targeting several kinases. In view of these facts, different computational approaches for kinase drug design need to be exploited, in order to find novel, more efficient and side effects-free kinase inhibitors.

Kinase sequence information had also been correlated with inhibition selectivity. A novel approach combines the understanding of small molecules and target sequence and genes, and thereby assists researchers in finding new targets for existing molecules or understanding selectivity and polypharmacology of molecules in related targets. Chemogenomics combines genomic data, structural biological data, classical dendrograms, and selectivity data to explore, define, and classify the medicinally relevant target space for any relevant biological system. Consequently, exploitation of this information in the discovery of kinase inhibitors defines practical kinase chemogenomics (kinomics) [13]. The authors presented the first dendrogram of kinases based entirely on small molecule selectivity data. They found that the selectivity dendrogram varied from sequence-based clustering due to the higher-level groupings of the smallest clusters, and it remains very comparable for closely homologous targets. As a main result, it was found that the smaller comparable molecules inhibit higher homologous kinases in a more desirable way.

In our study, the most relevant autocorrelation features were found to be thermodynamic transfer hydrophobicity (H_t), solvent-accessible reduction ratio (R_a) and solvent-accessibility area for native state ASA_N, which encode a kinases inhibitory pattern, defined by the distributions of hydrophobicity/polarity states along the sequence. At the same time, the differential affinity of ligands towards kinases was ruled by the distributions of atomic volume and electronegativity on the 2D structure sketches. To the best of our knowledge, our study is the first model for predicting inhibition data on 62 kinases and a wide chemical space, which allows discriminating between stable and unstable inhibition complexes with adequate accuracies about 82% for training set crossvalidation and test sets. The predictor is available online at: http://gibk21.bse.kyutech.ac.jp/AUTOkinI/SVMpredictor.html.

4.4. Conclusions

Kinase inhibition was successfully modeled from sequence and 2D graph representation of ligands using SVM. The topological model surpassed the fragment-based classifier with maximum crossvalidation accuracies about 82% for training set crossvalidation and test set prediction. The predictor was stable to the inclusion/exclusion of new kinases and accurately classified the affinity of totally new inhibitors in the test set. Furthermore, test set accuracies of the optimum topological classifier were very homogenous across 19 kinase families and 30 substructural fragments of the ligands.

4.5. References

1. Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34**:D204–D206.

2. Kim KH: **Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers?** *J Comput Aided Mol Des* 2007, **21**:421–435.

3. Woolfrey JR, Weston GS: **The use of computational methods in the discovery and design of kinase inhibitors**. *Curr Pharm Des* 2002, **8**:1527–1545.

4. Fernández L, Caballero J, Abreu JI, Fernández M: **Amino Acid Sequence Autocorrelation Vectors And Bayesian-Regularized Genetic Neural Networks For Modeling Protein Conformational Stability: Gene V Protein Mutants**. *Proteins* 2007, **67**:834–853.

5. Bogoyevitch MA, Fairlie DP: **A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding**. *Drug Discov Today* 2007, **12**:622–623.

6. Hochhaus A, La Rosee P: **Imatinib therapy in chronic myleogenous leukemia: Strategies to avoid and overcome resistance**. *Leukemia* 2004, **18**: 1321–1331.

7. Nagar B, Bornmann WG, Pellicena P, Schindler T, Veach DR, Miller WT, Clarkson B, Kuriyanet J: **Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571).** *Cancer Res* 2002, **62**:4236–4243.

8. Schindler T, Bornmann W, Pellicena P, Miller TW, Clarkson B, Kuriyan J: **Structural mechanism of STI-571 inhibition of Abelson tyrosine kinase**. *Science* 2000, **289**:1938–1942.

9. Manley WP, Cowan-Jacob WS, Mestan J: **Advances in the structural biology, design and clinical development of Bcr-Abl kinase inhibitors for the treatment of chronic myeloid leukaemia**. *Biochim Biophys Acta* 2005, **1754**:3–13.

10. Golas MJ, Arndt K, Etienne C, Lucas J, Nardin D, Boschelli D H, Boschelli F: **SKI-606, a 4-anilino-3-quinolinecarbonitrile dual inhibitor of Src and Abl kinases, is a potent antiproliferative agent against chronic myelogenous leukemia cells in culture and causes regression of K562 xenografts in nude mice**. *Cancer Res* 2003, **63**:375–381.

11. Lombardo LJ, Lee FY, Chen P, Norris D, Barrish JC, Behnia K, Castaneda S, Cornelius LA; Das J, Doweyko AM, Fairchild C, Hunt JT, Inigo I, Johnston K, Kamath A, Kan D, Klei H, Marathe P, Pang S, Peterson R, Pitt S, Schieven

GL, Schmidt RJ, Tokarski J, Wen ML, Wityak J, Borzilleri RM: **Discovery of N-(2-Chloro-6-methylphenyl)-2-(6-(4-(2-hydroxy- ethyl)-piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (BMS-354825), a dual Src/Abl kinase inhibitor with potent antitumor activity in preclinical assays**. *J Med Chem* 2004, **47**:6658–6661.

12. Thaimattam R, Banerjee R, Miglani R, Iqbal J: **Protein Kinase Inhibitors: Structural Insights Into Selectivity**. *Curr Pharm Des* 2007, **13**:2751–2765.

13. Vieth M, Higgs, RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H: **Kinomics-structural biology and chemogenomics of kinase inhibitors and targets**. *Biochim Biophys Acta* 2004, **1697**:243–257.

CHAPTER 5. MODELING OF PROTEASE INHIBITION

5.1. Introduction

This chapter presents the modeling of the ligand dataset with reported inhibitory activities towards 32 proteases collected from ProLiNT database [1]. Structural fragments and topological autocorrelations features were computed from protein sequences and 2D graphs of ligands. The dataset was divided into training (80% dataset) and test sets (20% dataset) by k-means clustering. Five clusters were generated and cases were homogeneously added to training and test sets by selecting instances from each cluster according to cluster's sizes. Subsequently, Support Vector Machines (SVMs) model was used to classify hypothetical complexes into high and low affinity classes. Predictor's optimization was carried out by three-fold-out (TFO) crossvalidation. The training set was divided into three subsets: two subsets were used for training the classifier and the rest subset was then predicted. This process was repeated until all the subsets were predicted. The ability of the predictor to recognize new ligands and targets was evaluated in the test set as well as the performance for different protease families and inhibitor scaffolds.

5.2. Results

5.2.1 Clustering analysis

Selectivity of the 9 protease families was firstly explored by cluster analysis. Two sets of descriptors, structural fragments and 2D autocorrelation vectors were calculated for the 988 active ligands on the dataset. Afterwards, k-means clustering algorithm yielded 9 ligand clusters (same as the number of protease families) for each descriptor sets. In order to evaluate the efficiency of proteases selectivity in the two descriptors sets, ligand clusters were matched to the different protease families.

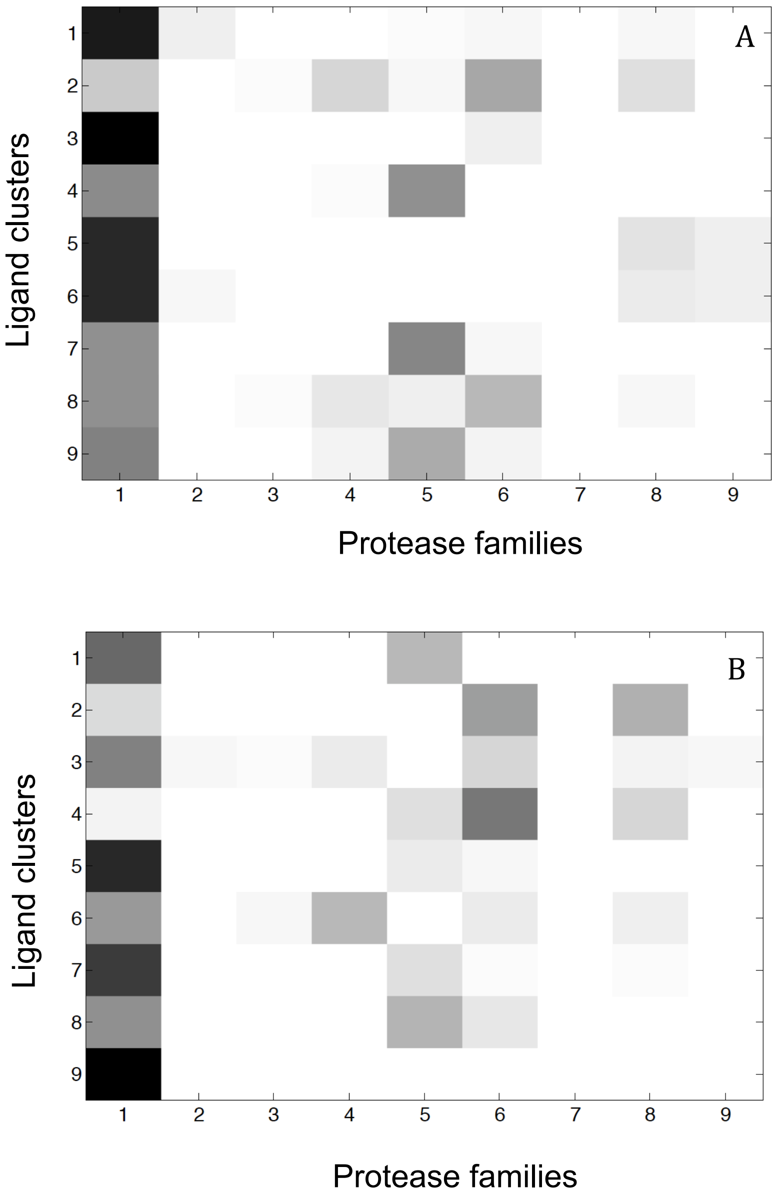
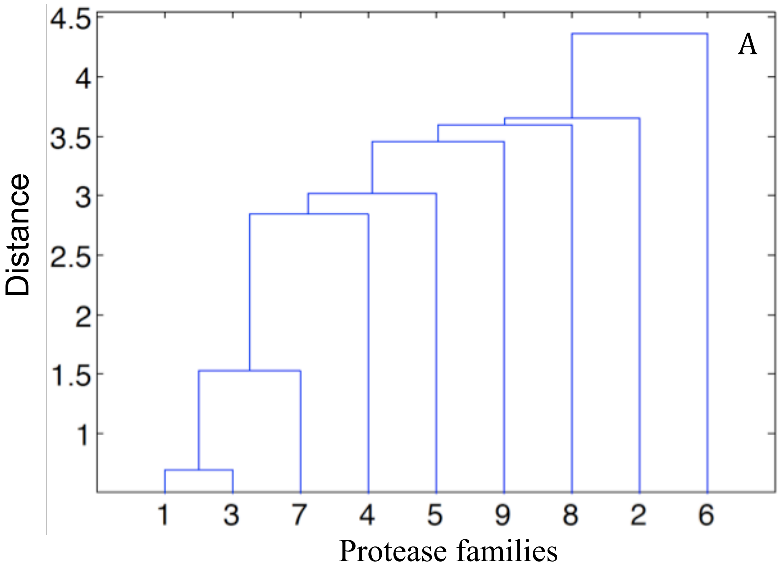


Figure 5.1. Density map representing occurrence ratios of protease families on active ligand clusters obtained by atom-center fragment descriptors (A) and 2D autocorrelation vectors (B). Protease families: (1) peptidase A2, (2) peptidase A1, (3) peptidase C1, (4) peptidase C2, (5) peptidase M10, (6) peptidase S1, (7) peptidase S9, (8) picornaviruses polyprotein, and (9) plasmepsin.

Distributions of both sets of 9 ligand clusters on the 9 protease families appear in Figure 5.1. The cluster distributions depict different patterns. Peptidase A2 family was the most populated one in both graphs but distributions of the ligand clusters on this protease family differ. In addition, we built two dendrograms by clustering protease families according to the distributions of the ligand clusters in the families. Figure 5.2 depicts dendrograms for fragment and topological descriptors showing 4 family clusters at squared Euclidean distances equal to 3.5 units in both cases but the distribution of protease families according each approach follow different trends.

The clustering pointed out that different chemical scaffolds inhibit independently allocated protease families. At the same time, the two families in multiple clusters could share similarity for the active ligands according to each encoding scheme. Among the highly represented families, allocation of peptidase A2 family and picornaviruses polyprotein family in the same cluster al low distance suggests a similar inhibition scaffold for these proteases according to the topological approach. The inhibition patterns of these families are similar and significantly differ from the rest of the protease families. Interestingly, HIV-1 protease in peptidase A2 family is a viral protease as well as the members of the picornaviruses polyprotein family.

Fragment and topological descriptors were evaluated to distinguish between active and inactive chemical scaffolds in the dataset. The mean squared differences between the normalized descriptors of active and inactive inhibitors were 0.0677 and 0.1878 from the fragment- and topological-based approaches, respectively. These values significantly differed from the mean squared differences of 100 scrambled partitions of active and inactive ligands that were $<10^{-4}$ for both descriptor types. Mean squared topological autocorrelation difference between active and inactive inhibitors is 3-fold higher in comparison with mean squared fragment difference.



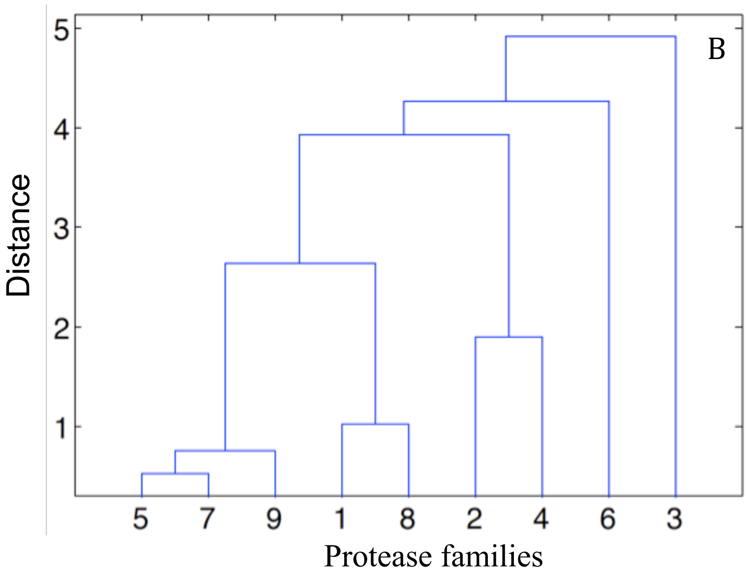


Figure 5.2. Dendrograms of protease families according to occurrences ratios on active ligand clusters obtained by fragment descriptors (A) and 2D autocorrelation vectors (B). Protease families: (1) peptidase A2, (2) peptidase A1, (3) peptidase C1, (4) peptidase C2, (5) peptidase M10, (6) peptidase S1, (7) peptidase S9, (8) picornaviruses polyprotein, and (9) plasmepsin.

5.2.2 Structural Fragments classifier

SF and *SSF* descriptors computed for the protease sequences and the 2D structural sketches of the inhibitors were combined into a single feature matrix by simple concatenation of target and ligand descriptors blocks. The dataset was separated into a training set with 1279 inhibition complexes (80%) and a test set with 427 inhibition complexes (20%).

Table 1. Crossvalidation and prediction statistics for the training and test sets according to the structural fragment SVM model for the classification of the stability of protease inhibition complexes. SVM regularization parameter and width of the RBF kernel were 2 and 5, respectively.

Experiment	***Q2***	***P(+)***	***P(-)***	***Q(+)***	***Q(-)***	***Cr***
Training set crossvalidation	0.75	0.73	0.79	0.75	0.73	0.51
Test set prediction	0.74	0.78	0.70	0.83	0.72	0.50

+ and - : the indexes were evaluated for “unstable” (*Ki* < 0.1 μM) and “stable” (*Ki* > 0.1 μM) protease inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *P(s)* is the number of correct prediction for class s/all prediction made for s; *Q(s)* is the number of correct prediction for class s/observed in class s; *Cr* is Matthews’s correlation coefficient. *Q(+)* and *Q(-)* are sensitivity and specificity of high affinity class prediction and *P(+)* and *P(-)* are precision scores.

In a first attempt, we implemented a linear kernel but the crossvalidation accuracy was < 70%. Then, a nonlinear SVM classifier was optimized by adjusting RBF kernel width and regularization parameter throughout a TFO crossvalidation in a grid search. Optimum values of RBF kernel width and regularization parameter were 5 and 2, respectively, yielding crossvalidation results that appear in Table 5.1. An overall TFO crossvalidation accuracy of 75% for the classification of inhibition complexes was achieved with a correlation coefficient *C*=0.51.

5.2.3 Topological autocorrelation classifier

Topological features were computed from the primary sequence of the proteases and the 2D sketch of the inhibitor structure descriptors. Similar to the fragment-based classifier, the dataset was separated into training and test sets in the same way.

Table 5.2. Crossvalidation and prediction statistics for training and test sets according to the topological SVM model for the classification of the stability of protease inhibition complexes. SVM regularization parameter and width of the RBF kernel were 10 and 0.10, respectively.

Experiment		***Q2***	***Q(+)***	***Q(-)***	***P(+)***	***P(-)***	***Cr***
Training set crossvalidation	Complex-wise	0.81	0.81	0.81	0.85	0.76	0.61
	Ligand-wise	0.79	0.79	0.79	0.83	0.74	0.58
	Protease-wise	0.75	0.71	0.79	0.82	0.68	0.50
Test set prediction	Complex-wise	0.80	0.81	0.78	0.85	0.72	0.58
	Ligand-wise	0.83	0.86	0.73	0.90	0.64	0.57

+ and - : the indexes were evaluated for "low affinity" ($Ki > 0.1$ μM) and "high affinity" ($Ki < 0.1$ μM) protease inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *P(s)* is the number of correct prediction for class s/all prediction made for s; *Q(s)* is the number of correct prediction for class s/observed in class s; *Cr* is Matthews's correlation coefficient. *Q(+)* and *Q(-)* are sensitivity and specificity of high affinity class prediction and *P(+)* and *P(-)* are precision scores.

A linear kernel explained < 70% of training set crossvalidation variance. Then, adjusting RBF kernel and SVM parameter throughout a TFO crossvalidation in a grid search yielded a nonlinear classifier with crossvalidation results shown in Table 5.2. Here complex-wise statistics refers to the basic crossvalidation experiments in which only nonredundant complexes were included. In this case, an overall complex-wise TFO crossvalidation accuracy of 81% was achieved for the classification of inhibition complexes with a correlation coefficient *C*=0.65. The Crossvalidation accuracies to identify high and low affinity inhibition complexes were also 81%. These accuracies and the correlation coefficient are higher than the statistics reported in Table 5.1 for the fragment-based predictor. In the view of the training set results, which are similar to our findings for kinase inhibition, we conclude that autocorrelation approach outperforms the fragment-based classifier and it is more convenient for modeling proteases inhibition.

Moreover, the stability of the optimum topological model to recognize new ligands and proteases was evaluated. We performed two additional crossvalidation experiments in which protease complexes sharing similar ligands were kept in the same data subset during crossvalidation, we called this experiments ligand-wise crossvalidation, similarly we run another crossvalidation in which complexes of the same protease were kept in the same subset and this was called protease-wise crossvalidation. The accuracies for ligand-wise crossvalidation were 75% (Table 5.2) illustrating that the model correctly classifies the affinity of novel ligands towards existing protease. All further reported statistical analysis for protease families (Figures 5.3 and 5.4) and ligand chemotypes (Table 5.1 Appendix) were performed on ligand-wise crossvalidation. Noteworthy, the protease-wise TFO crossvalidation in Table 5.2 showed how the model differentiated complexes of new proteases with accuracy of 75% and recognized low and high affinity complexes with accuracies of 79% and 71%. This result is remarkable considering that removing highly represented proteases from the training subset leaves out large series of inhibitors. This topological SVM predictor of protease inhibition outperforms our previous model on kinases in detecting new high affinity targets. In fact, the protease inhibition depicts a more homogenous pattern across different targets than the kinase inhibition does. Ligand- and target-wise crossvalidations strongly corroborate the relevance of the topological feature space to model protease inhibition as well as the self-consistency of the SVM predictor.

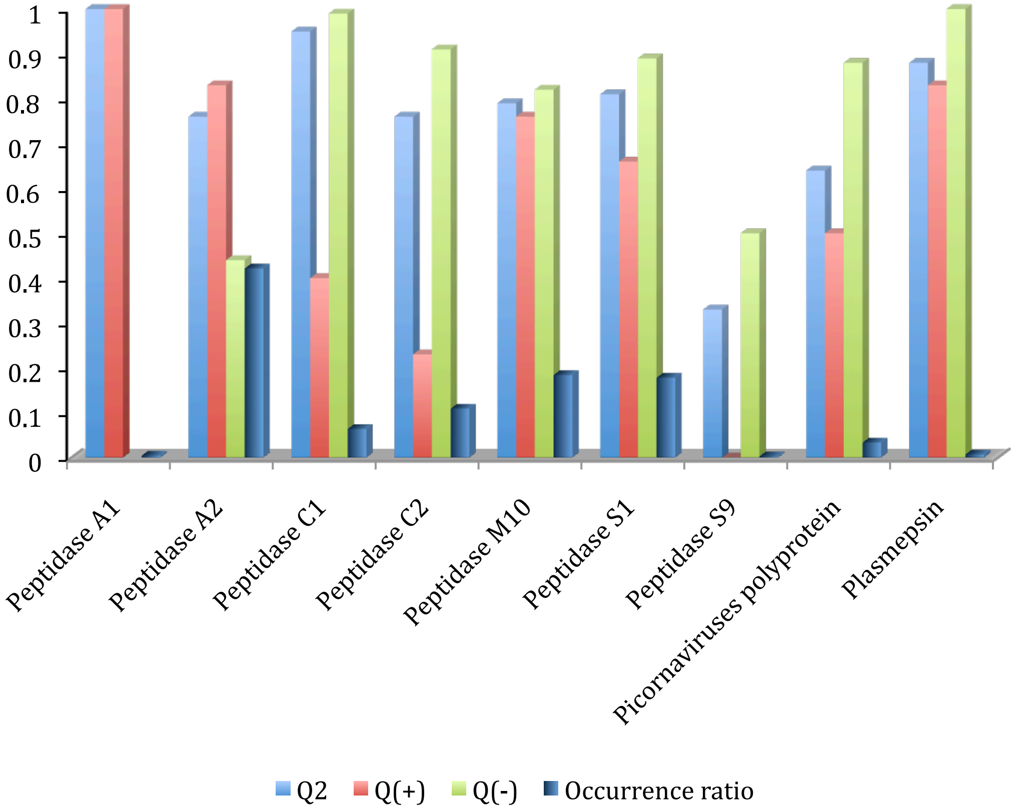


Figure 5.3. Ligand-wise TFO crossvalidation accuracies for the 9 protease families in the training set according to the optimum topological SVM model. + and - : the indexes were evaluated for "low affinity" (*Ki* > 0.1 μM) and "high affinity" (*Ki* < 0.1 μM) protease inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of high affinity class prediction.

5.2.4 Performance of the topological autocorrelation classifier on different protease families and chemotypes on the training set

Protease inhibition dataset includes inhibitory activities of a diverse chemical space towards 9 protease families. It is very interesting to analyze the optimum classifier performance for each protease family in the dataset. The classification accuracies for each family according to the optimum SVM predictor are depicted in Figure 5.3. The predictor performance was homogeneous to all families. The overall accuracies for the classification of the affinity of inhibition complexes were higher than 65% for all proteases with the exception of peptidase S9 family. Interestingly, the same protease family was found isolated in the cluster analysis. Therefore, the inhibition pattern of this protease group is more complex and largely differs from the others. This family is integrated by proline-specific peptidases (most of them serine-dependent peptidases), which have been found in bacteria, protozoa, plants, and animals, including mammals. Nevertheless, these enzymes do not seem to share similar functions across these organisms [2, 3]. The classifier recognized low affinity inhibition complexes of the peptidase A2 proteases with low accuracy. This family included HIV-1 protease and the poor result in crossvalidation experiment could be caused by the complexity and variety of the target-ligand interactions for this enzyme [4].

According to Figure 5.3, protease families with the highest occurrence ratios on the training set exhibit crossvalidation accuracies > 75%. These protease families are peptidase A2, peptidase M10, peptidase S1, peptidase C2 and peptidase C1 families with occurrence ratios of 42.14%, 18.37%, 17.75%, 10.87% and 6.33%, respectively. Peptidase M10 family have been extensively studied for targeting cancer. They are involved in connective-tissue remodeling and participate in some processes such as ovulation, embryonic growth, angiogenesis, differentiation, and healing [5]. MMP inhibitors have caught the interest as an important class of drugs for the development of innovative chemotherapeutics in several fields where effective treatments are lacking [5]. Considering the relevance of this protease family as therapeutic target, it is noteworthy that our classifier successfully discriminated between high and low affinity inhibitors of MMP proteases with an overall accuracy > 80% in crossvalidation experiments.

Cystein proteases belong to the peptidase C1 and peptidase C2 families in which a catalytic cysteine mediates protein hydrolysis via nucleophilic attack on the carbonyl carbon of a susceptible peptide bond [6]. Cysteine proteases have been identified as promising targets for the development of antiparasitic chemotherapy [7]. Their widespread importance in both protozoan and helminth parasites of domestic animals and humans along with the relative lack of redundancy of cysteine proteases in parasites compared to their mammalian hosts makes them attractive targets for the development of new antiparasitic chemotherapy [7]. Clinical studies have confirmed the efficacy of cysteine protease inhibitors in treatment of *Trypanosoma cruzi*, *Plasmodium falciparum*, *Leishmania major* and *Trypanoasoma brucei*, the agent of African trypanosomiasis. Considering this, the topological model that recognizes inhibition complexes of peptidase C1 proteases with crossvalidation accuracy > 90% could make a good contribution to the quest for potent inhibitors.

The second most abundant protease family in our data set, peptidase S1 family, is the import group of serine endopeptidases. Serine proteases perform many relevant functions in multi-cellular organisms, including blood clotting, cancer, cell death, osmoregulation, tissue remodeling, and immunity to infection [8-10]. Serpins are serine protease inhibitors that play a fundamental role in controlling the recognition of antigen, effector function, and homeostatic control of cytotoxic T lymphocytes through the inhibition of physiological serine protease targets [8]. In turn, serine proteases such as kallikrein, thrombin, and plasmin mediate the decrease of blood loss and transfusion requirements during coronary artery bypass graft [9]. They also play a critical role in neuronal death after injury to the central nervous system [10]. The neuroprotective activity of the molecules that modulate the activity of serine proteases have encouraged the study of the production and modulation of serpins, such as type 1 plasminogen activator inhibitor, neuroserpin, and protease nexin-1 [10]. The significance of serine proteases as therapeutic targets has boosted the development of synthetic inhibitors by combining experimental and computational medicinal chemistry [11-16]. In this context, our novel predictor is a valuable tool that correctly recognizes about 70% and 90% of active and inactive inhibitors towards serine proteases in crossvalidation experiments.

The clinical efficacy of viral proteases inhibitors has been demonstrated in HIV infection. Peptidase A2 family includes HIV-1 protease, which is structurally related to human aspartic acid proteases, such as renin, cathepsin D, gastrin, and pepsin [17]. The earliest HIV protease inhibitors were peptidic in nature with limited success due to their poor oral bioavailability and metabolic instability. Subsequently, the screening of compound libraries and rational drug design based on solved X-ray crystal structures of HIV-1 protease have led to many good inhibitors, such as nelfinavir [18]. Although, some HIV protease inhibitors have emerged as potent antiretroviral chemotherapeutic agents, the emergence of drug-resistant variants after prolonged antiviral therapy leads to resistance to one protease inhibitor and cross-resistance to other protease inhibitors. This scenario demands continuous drug design efforts, which is likely to be supported by our study as in this case our model has an accuracy of 76%.

We also analyzed the predictor behavior for different chemical subspaces on the inhibitor dataset. In this sense, 25 substructural templates were considered for comparing the classifier accuracy regarding the different chemotypes on the modeled chemical space (Table 5.1 Appendix). All but two of the analyzed substructures showed overall accuracies about or higher than 70%, where ligands bearing bromobenzene or thiofuran were predicted with overall accuracies about 50%. The accuracies for separate classes were lower than 50% only for low affinity ligands bearing 1,3-dichlorobenzene substructure and high affinity ligands bearing thiofuran. According to this result, we can state that the classifier performed well over the chemical space represented by the protease inhibitors in the training set.

5.2.4 Prediction of the test set

Crossvalidation accuracy gives an estimate of the internal consistency of the predictive models, but a more realistic measurement of the prediction power can be achieved by evaluating performance on blind test set. Table 5.2 shows that the results from the topological-based classifier on such datasets perform well and lead to an overall accuracy of about 80% for the complex-wise evaluation. This is promising considering the fact that the test set prediction accuracies were in the same range as obtained in crossvalidation experiments of the training set in Table 5.2, thus relaxing the concerns of overfitting. Classification of test set performed well for the topological models taking into account that test set accuracy of the fragment-based predictor was 75% (Table 5.1).

In addition, the ability of the topological predictor to recognize totally new ligands was estimated by evaluating the model accuracy for 248 out of the 427 protease inhibition complexes in the test set for which ligands information was not available in the training set. This predictor correctly recognized 83% of totally new high affinity ligands with an adequate specificity of 73%. This result showed that the optimum classifier not only correctly learned the protease inhibition pattern, but that the learned pattern was adequately generalized to the test set, including totally new ligands.

Figure 5.4 shows classification results for each protease family on the test set. Peptidase S9 family was totally misclassified but this protease family is not statistically significant in the test set with the lower occurrence ratio of 0.23%. In turn, the rest of the protease families, with the exception of picornaviruses polyprotein family, exhibit overall prediction accuracies over 65%. Furthermore, the classifier predicts five out of nine protease families in the test set with accuracies about 85%.

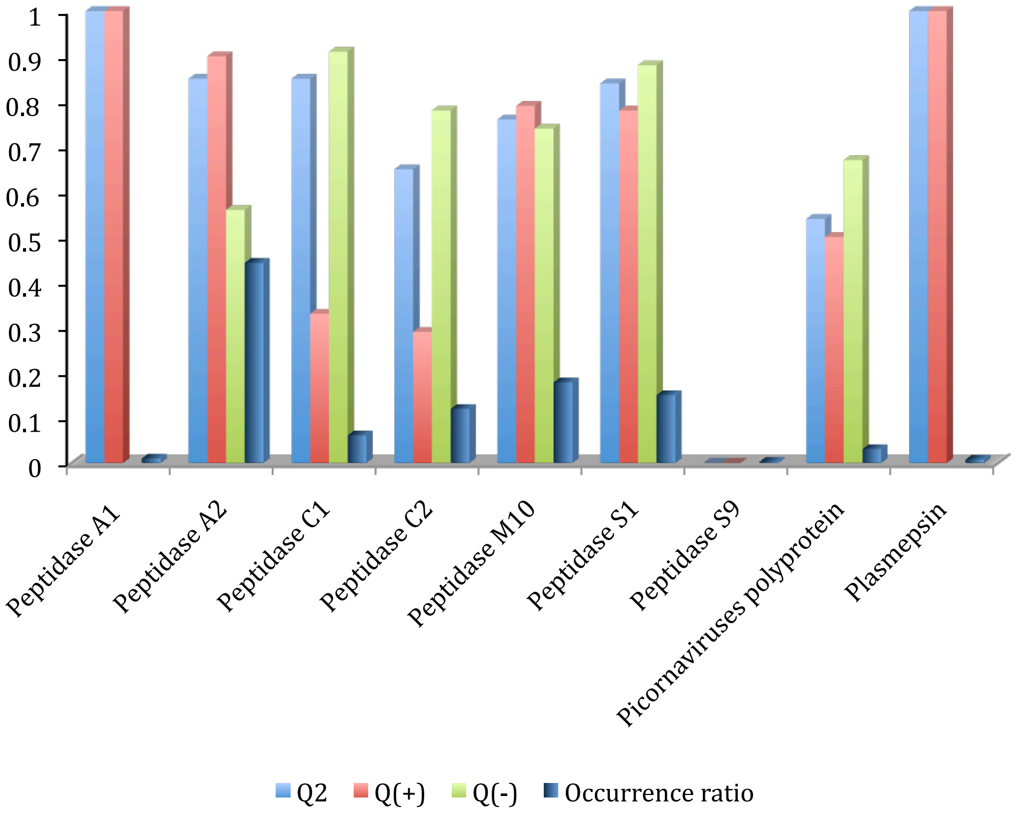


Figure 5.4. Ligand-wise prediction accuracies for the 9 protease families in the test set according to the optimum topological SVM model. + and - : the indexes were evaluated for "low affinity" (*Ki* > 0.1 µM) and "high affinity" (*Ki* < 0.1 µM) protease inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of high affinity class prediction.

The classifier performance on the test set according to the different chemotypes was similar to the crossvalidation of the training set (Table 5.1 in Appendix). All inhibitor chemotypes have overall accuracies > 50% but 1,3-difluorobenzene and biphenyl have overall accuracies < 75%. Low affinity inhibitors bearing tetramethylmethane, biphenyl, chlorobenzene and indole chemotypes were recognized with low accuracies. Even though the low prediction performance for these chemotypes, the overall performance of the predictor on this blind set is still encouraging. The classifier recognized the inhibition pattern from different protease families and accurately discriminated between high and low affinity inhibitor complexes belonging to several chemical subspaces.

The differential relevance of the topological autocorrelation vectors for modeling protease inhibition was evaluated by sensitivity analysis [19]. The impact of each variable in the optimum model was estimated by measuring test set prediction accuracies for modified feature matrices in which each autocorrelation vector at a time was replaced by a constant vector of same length. The magnitude of the importance of each input variable in the predictor was taken as the underperformance score defined as the ratio between *F-scores* (Eq. 2.11) for original and modified feature matrices. The sensibility analysis yields top-14 relevant *AASA* vectors with approximately equal contributions to the classification model as follows: from $AASA3H_t$ to $AASA5H_t$, from $AASA1pK'$ to $AASA5pK'$, from $AASA1R_a$ to $AASA5R_a$ and $AASA3ASA_N$. The top-5 relevant inhibitor autocorrelation vectors in descending order were as follows: $ATS8v > ATS4e > ATS7v > ATS8e > ATS8p$. It is noteworthy that the most relevant inputs are protease's autocorrelations of hydrophobicity/polarity-related properties such as thermodynamic transfer hydrophobicity (H_t), solvent-accessible reduction ratio (R_a), solvent-accessibility area for native state (ASA_N) and the electrostatic state related property equilibrium constant with reference to the ionization property of COOH group (pK'). These properties combined with ligand's autocorrelations of atomic volumes, electronegativities and polarizabilities at middle to large lags on the 2D structure sketch depict a robust protease inhibition pattern.

5.3. Discussion

Initially, protease inhibitors were developed by natural product screening for lead compounds followed by empirical substrate-based optimization [20]. The availability of three-dimensional structural information for proteases later improved this substrate-based drug design allowing receptor-based computational design. Classical and 3D-QSAR approaches have been extensively used to model HIV-1 protease inhibition [21-23]. Furthermore, machine learning techniques have been successfully implemented to predict HIV protease inhibitor activities yielding accuracies in the range 70%-80% [24, 25]. The inhibition of gamma-secretase, a multi-subunit protease complex of the aspartyl protease family and a promising target for the prevention and treatment of Alzheimer's disease, was modeled by random forest yielding classification accuracy of 99% and about 350 potential hit candidates in a broad virtual screening against the ZINC database [26].

The design of selective MMP inhibitors has been a difficult task because members of the MMP family share structural features including propeptide, catalytic, and hemopexin domains. Neural networks models for individual members of this enzyme family have been reported with accuracies in the range 72%-81%. Such studies partially succeed in distinguishing the structural requirements for the selective inhibitors of the MMP proteases but did not consider target information. High prediction performance was reported for an external test set of more than 100 compounds using ligand alignments from X-ray structures and 3D-QSAR to identify MMP3 inhibitors [29]. Furthermore, the in vitro evaluation of active candidate yielded some nanomolar active inhibitors [29]. Verma and Hansh [30] profoundly analyzed MMP family from the classification of these enzymes to the clinical trials of their inhibitors. About a hundred of published and newly formulated QSAR models on MMP inhibition were discussed in the context of the chemical–biological interactions. The most important features were hydrophobicity and molar refractivity. The target information was not explicitly used but the concluded target-ligand interaction hypothesis involved comparing the quality and descriptor occurrences on the models for different inhibitor datasets on the same target or the same inhibitor dataset for different targets. Despite that rigor in this area, the authors used very intuitive descriptors and, although the stability of target-ligands were predicted with high crossvalidation accuracies, the use and generalization of about hundred models as well as their comparative interpretation is rather rough.

In turn, the inhibition of serine protease thrombin was modeled by a method called Linear Interaction Energy in Continuum Electrostatics, which successfully predicted the inhibitory activity of about 76% of the compounds in an external test set. Active compounds were selected with high sensitivity from a thrombin combinatorial library of more than 10000 mimetic chemicals [12]. Similarly, a group of low molecular weight cathepsin B inhibitors were predicted by partial least-squares (PLS) QSAR models with regression and classification accuracies of 68% and 94% in crossvalidation experiments [13]. In another study, the 3D structure of the receptor complex of protease NS3, a serine protease that participates in the replication and maturation of Dengue virus [14], was used to computationally design and screen a virtual library of about 1000 peptidomimetic analogs [15]. The most promising virtual hits had inhibition potencies in nanomolar range and Adsorption, Distribution, Metabolism and Excretion (ADME) ADME-related properties comparable to the training set inhibitors [34]. An interesting review on FXa inhibitors discussed 3D-QSAR studies and classical QSAR approaches on chemically diverse data sets ranging from 20 to 80 compounds [16]. The most correlating feature was hydrophilicity-related properties as ClogP and molar refractivity and sterimol parameters were also important in most of the cases.

Recent QSAR analyses on the major cysteine protease from *Trypanosoma cruzi*, cruzain highlighted this enzyme as a very attractive drug target. Neural network [31] and 3D-QSAR [32] studies of cruzain inhibitors described 73-78% of crossvalidation data variances and the predicted values were by experimental results.

Although protease inhibition have been extensively approached by computational methods, the existing predictors lack a comprehensive and unified implementation that integrates wide targets and ligands spaces. In our study, we employed ligand clustering to evaluate the ability of ligand cluster distributions to differentiate the protease families. In addition, optimum modeling of protein-ligand interactions was developed from simple sequence and 2D graphs combining topological descriptors of targets and ligands. Autocorrelation vectors weighted by amino acids/residues and atomic properties encode target sequences and inhibitor structures. Those descriptors account for amino-acid distributions on the target sequences and atom distribution on the 2D sketch of the inhibitor molecules. While combining topological autocorrelation features, interactions between target and ligand structures are encoded in a conformation-independent set of descriptors. The most relevant autocorrelation features were found to be thermodynamic transfer hydrophobicity (H_t), solvent-accessible reduction ratio (R_a), the electrostatic state related property equilibrium constant with reference to the ionization property of COOH group (pK') and solvent-accessibility area for native state ASA_N, which encode a protease inhibitory pattern, defined by the distributions of hydrophobicity/polarity and electronegativity states along the target sequence. At the same time, the differential affinity of ligands towards proteases was ruled by the distributions of atomic volume, electronegativity and polarizability on the 2D structure sketches. To the best of our knowledge, our study is the first model for predicting inhibition data on 32 proteases and a wide chemical space, which allows discriminating between high and low affinity inhibition complexes with adequate accuracies > 80% for the test sets. The predictor is available online at:

http://gibk21.bse.kyutech.ac.jp/AUTOprotI/SVMpredictor.html.

5.4. Conclusions

Protease inhibition was successfully modeled from sequence and 2D graph representation of ligands using SVMs. A topological model using sequence information surpassed a fragment-based classifier with maximum crossvalidation accuracies >80% for training set crossvalidation and test set .The predictor was stable to the inclusion/exclusion of new proteases and accurately classified the affinity of totally new inhibitors in the test set. Furthermore, test set accuracies of the optimum topological classifier were homogenous across the 32 protease families and 25 substructural fragments of the ligands.

5.5. References

1. Ahmad S, Kitajima K, Selvaraj S, Kubodera H, Sunada S, An J-H, Sarai A: **Protein-Ligand Interactions: ProLINT Database and QSAR Analysis** *Genome Inform* 2003, **14**:537–538.

2. Rosenblum JS, Kozarich JW: **Prolyl peptidases: a serine protease subfamily with high potential for drug discovery**. *Curr Opin Chem Biol* 2003, **7**:496–504.

3. Polgar L: **The prolyl oligopeptidase family**. *Cell Mol Life Sci* 2002, **59**:349–362.

4. Leung D, Abbenante G, Fairlie DP: **Protease Inhibitors: Current Status and Future Prospects**. *J Med Chem* 2000, **43**:305–341.

5. Baker AH, Edwards DR, Murphy G: **Metalloproteinase inhibitors: biological actions and therapeutic opportunities**. *J Cell Sci* 2002, **115**:3719–3727.

6. Rawlings ND, Tolle DP, Barrett AJ: **Evolutionary families of peptidase inhibitors.** *Biochem J* 2004, **378**:705–716.

7. McKerrow JH, Engel JC, Caffrey CR. **Cysteine protease inhibitors as chemotherapy for parasitic infections**. *Bioorg Med Chem* 1999, **4**:639–644.

8. Ashton-Rickardt PG: **Serine protease inhibitors and cytotoxic T lymphocytes.** *Immunol Rev* 2010, **235**:147–158.

9. Nicolotti O, Fabiola T, Andrea M, Leonetti CF, Carotti A: **An Integrated Approach to Ligand- and Structure-Based Drug Design: Development and Application to a Series of Serine Protease Inhibitors.** *J Chem Inf Model* 2008, **48**:1211–1226.

10. Walker B, Lynas JF: **Strategies for the inhibition of serine proteases.** *Cell Mol Life Sci* 2001, **58**:596–624.

11. Perekhodtsev GD: **Similarity study of serine proteases inhibitors**. *Mol Diver* 2006, **10**:81–83.

12. Nicolotti O, Giangreco I, Miscioscia TF, Convertino M, Leonetti F, Pisani L, Carotti A: **Screening of benzamidine-based thrombin inhibitors via a linear interaction energy in continuum electrostatics model**. *J Comput Aided Mol Des* 2010, **24**:117–129.

13. Zhou Z, Wang Y, Bryant SH: **QSAR models for predicting cathepsin B inhibition by small molecules–continuous and binary QSAR models to classify cathepsin B inhibition activities of small molecules**. *J Mol Graph Model* 2010, 28:714–27.

14. Chambers TJ, Weir RC, Grakoui A, McCourt DW, Bazan JF, Fletterick RJ,

Rice CM: **Evidence that the N-terminal domain of nonstructural protein NS3 from yellow fever virus is a serine protease responsible for site-specific cleavages in the viral polyprotein**. *Proc Natl Acad Sci USA* 1990, **87**:8898–8902

15. Frecer V,Miertus S: **Design, structure-based focusing and in silico screening of combinatorial library of peptidomimetic inhibitors of Dengue virus NS2B-NS3 protease**. *J Comput Aided Mol Des* 2010, 24:195–212.

16. Kontogiorgis CA, Hadjipavlou-Litina, D: **Current Trends in Quantitative Structure Activity Relationships on FXa inhibitors: Evaluation and Comparative Analysis**. *Med Res Rev* 2004, **24**:687–747.

17. Von der Helm K: **Retroviral proteases: structure, function and inhibition from a non-anticipated viral enzyme to the target of a most promising HIV therapy**. *Biol Chem* 1996, **377**:765–774.

18. Patick AK, Potts KE: **Protease Inhibitors as Antiviral Agents.** *Clin Microbiol Rev* 1998, **11**:614–627.

19. Fernández L, Caballero J, Abreu JI, Fernández M: **Amino Acid Sequence Autocorrelation Vectors And Bayesian-Regularized Genetic Neural Networks For Modeling Protein Conformational Stability: Gene V Protein Mutants**. *Proteins* 2007, **67**:834–853.

20. West ML, Fairlie DP: **Targeting HIV-1 protease: a test of drug-design methodologies**. *Trends Pharmacol Sci* 1995, **16**:67–75.

21. Gupta SP, Babu MS, Sowmya S: **A Quantitative Structure-Activity Relationship Study on Some Sulfolanes and Arylthiomethanes Acting as HIV-1 Protease Inhibitors**. *Bioorg Med Chem* 1998, **6**:2185–2192.

22. Huang X, Xu L, Luo X, Fan K, Ji R, Pei G, Chen K, Jiang H: **Elucidating the Inhibiting Mode of AHPBA Derivatives against HIV-1 Protease and Building Predictive 3D-QSAR Models**. *J Med Chem* 2002, **45**:333–343.

23. Katritzky AR, Oliferenko A, Lomaka A, Karelson M: **Six-Membered Cyclic Ureas as HIV-1 Protease Inhibitors: A QSAR Study Based on CODESSA PRO Approach**. *Bioorg Med Chem Lett* 2002, **12**:3453–3457.

24. Patankar SJ, Jurs PC: **Classification of HIV protease inhibitors on the basis of their antiviral potency using radial basis function neural networks**. *J Comput Aided Mol Des* 2003, **17**:155–71.

25. Fernández M, Caballero J: **Modeling of Activity of Cyclic Urea HIV-1 Protease Inhibitors using Regularized-Artificial Neural Networks**. *Bioorg Med Chem* 2006, **14**:280–294.

26. Yang XG, Lv W, Chen YZ, Xue Y: **In silico prediction and screening of gamma-secretase inhibitors by molecular descriptors and machine learning methods**. *J Comput Chem* 2010, **31**:1249–1258.

27. Fernández M, Caballero J, Tundidor-Camba A: **Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino] acetamide derivatives as matrix metalloproteinase inhibitors**. *Bioorg Med Chem* 2006, **14**:4137–4150.

28. Fernández M, Caballero J: **QSAR Modeling of matrix metalloproteinase inhibition by N-Hydroxy-α-henylsulfonylacetamide derivatives**. *Bioorg Med Chem* 2007, **15**:6298–6310.

29. Tuccinardi T, Ortore G, Santos MA, Marques SM, Nuti E, Rossello A, Martinelli A: **Multitemplate alignment method for the development of a reliable 3D-QSAR model for the analysis of MMP3 inhibitors**. *J Chem Inf Model* 2009, **49**:1715–1724.

30. Verma RP, Hansch C: **Matrix metalloproteinases (MMPs): Chemical–biological functions and (Q)SARs**. *Bioorg Med Chem* 2007, **15**:2223–2268.

31. Caballero J,Tundidor-Camba A,Fernández M: **Modeling of the inhibition constant (*Ki*) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized Genetic Neural Networks**. *QSAR Comb Sci* 2007, **26**:27–40.

32. Trossini GH, Guido RV, Oliva G, Ferreira EI, Andricopulo AD: **Quantitative structure-activity relationships for a series of inhibitors of cruzain from Trypanosoma cruzi: molecular modeling CoMFA and CoMSIA studies**. *J Mol Graph Model* 2009, **28**:3–11.

CHAPTER 6. SUMMARY AND FUTURE PERSPECTIVES

6.1. Summary

The present study investigated the modeling of protein properties and function combining simple representation of the structure and machine learning techniques. Three datasets of protein conformational stability, kinase and proteases inhibition were collected from our in-house databases Protherm and ProLINT. Proteins and ligands were represented as Cα-carbon linear graphs and 2D molecular graphs from which autocorrelation features were computed using different weighting schemes. Support vector machines were implemented to map autocorrelation features to the property/function. Feature selection was applied for protein conformational stability. The main results are the following:

1. **Conformational stability**
   - Real *ΔΔ*G values were calculated with correlation coefficient of 0.67.
   - *ΔΔ*G signs were correctly classified with accuracy of 80%.
   - Mutations at the protein core were more accurate predicted in comparison to mutations at the protein surface.
   - Protein hydration-related properties rule the prediction of conformational stability.
2. **Kinase inhibition**
   - AGC Ser/Thr protein kinase, Tyr protein kinase and atypical kinase families had exclusive inhibition profiles according to the fragment and autocorrelation patterns of active inhibitors.
   - Overall inhibition prediction has accuracy > 80%.
   - 15 out of 19 kinase families have accuracy of 85%.

- Inhibition of TKL Ser/Thr protein kinase family was predicted with low accuracy suggesting a complex ligand-target interaction profile.
- Most relevant features resemble the hydrophobicity/polarity states along the kinase sequences.

3. **Protease inhibition**

- Peptidase A2 and virus picornaviruses polyprotein families have similar and exclusive inhibition profiles according to autocorrelation pattern of active inhibitors.
- Overall inhibition prediction has accuracy > 80%.
- Five out of nine protease families have accuracy of 85%.
- Inhibition of peptidase A2 and peptidase S9 families were predicted with low accuracy suggesting complex ligand-target interaction profiles.
- Most relevant features resemble the hydrophobicity/polarity and electronegativity states along the protease sequences.

6.2. Future perspective

This study focuses on the prediction of protein conformational stability and kinase and protease inhibition using sequences and 2D molecular graphs. It is an attempt to improve the tools to analyze and query structure databases. These results show that accurate predictor can be implemented from a proper representation of the scare information content in protein sequence and 2D molecular graphs. It provides efficient and robust prediction tools to screen sequence and ligand databases for putative stable or unstable mutants and potent kinase and proteases inhibitors. Despite high to moderate accuracies were obtained; in some cases the complexity of the interactions limited the ability of the machine learning technique to reproduce the learned information. In these particular cases, tridimensional structure information, added to the models in combination with solvent accessibility, secondary structure and evolutionary data, would further improve performance.

Appendix

Table 2.1 Appendix. Numerical values of 48 selected physicochemical, energetic, and conformational properties of the 20 amino acids/residues used in protein conformational stability modeling.

Property^{a,b}		A	C	D	E	F	G	H	I	K
1	K_0	-25.5	-32.82	-33.12	-36.17	-34.54	-27	-31.84	-31.78	-32.4
2	H_t	0.87	1.52	0.66	0.67	2.87	0.1	0.87	3.15	1.64
3	H_P	13.05	14.3	11.1	11.41	13.89	12.2	12.42	15.34	11.01
4	P	0	1.48	49.7	49.9	0.35	0	51.6	0.1	49.5
5	pH_i	6	5.05	2.77	5.22	5.48	5.97	7.59	6.02	9.74
6	pK'	2.34	1.65	2.01	2.19	1.89	2.34	1.82	1.36	2.18
7	M_w	89	121	133	147	165	75	155	131	146
8	P_l	11.5	13.46	11.68	13.57	19.8	3.4	13.67	21.4	15.71
9	R_f	9.9	2.8	2.8	3.2	18.8	5.6	8.2	17.1	3.5
10	m	14.34	35.77	12	17.26	29.4	0	21.81	19.06	21.29
11	H_{nc}	0.62	0.29	0.9	-0.74	1.19	0.48	-0.4	1.38	-1.5
12	E_{sm}	1.4	1.37	1.16	1.16	1.14	1.36	1.22	1.19	1.07
13	E_l	0.49	0.67	0.35	0.37	0.72	0.53	0.54	0.76	0.3
14	E_t	1.9	2.04	1.52	1.54	1.86	1.9	1.76	1.95	1.37
15	P_a	1.42	0.7	1.01	1.51	1.13	0.57	1	1.08	1.16
16	P_b	0.83	1.19	0.54	0.37	1.38	0.75	0.87	1.6	0.74
17	P_t	0.66	1.19	1.46	0.74	0.6	1.56	0.95	0.47	1.01
18	P_C	0.71	1.19	1.21	0.84	0.71	1.52	1.07	0.66	0.99
19	C_a	20	25	26	33	46	13	37	39	46
20	F	0.96	0.87	1.14	1.07	0.69	1.16	0.8	0.76	1.14
21	P_r	0.38	0.57	0.14	0.09	0.51	0.38	0.31	0.56	0.04
22	R_a	3.7	3.03	2.6	3.3	6.6	3.13	3.57	7.69	1.79
23	N_s	6.05	7.86	4.95	5.1	6.62	6.16	5.8	7.51	4.88
24	a_n	1.59	0.33	0.53	1.45	1.14	0.53	0.89	1.22	1.13
25	a_c	1.44	0.76	2.13	2.01	1.01	0.62	0.56	0.68	0.59
26	a_m	1.22	1.53	0.56	1.28	1.13	0.4	2.23	0.77	1.65
27	V^0	60.46	67.7	73.83	85.88	121.48	43.25	98.79	107.72	108.5
28	N_m	2.11	1.88	1.8	2.09	1.98	1.53	1.98	1.77	1.96
29	N_l	3.92	5.55	2.85	2.72	4.53	4.31	3.77	5.58	2.79
30	H_{gm}	13.85	15.37	11.61	11.38	13.93	13.34	13.82	15.28	11.58
31	ASA_D	104	132.5	132.2	161.9	182	73.4	165.8	171.5	195.2
32	ASA_N	33.2	17.9	62.4	81	33.1	29.2	57.7	28.3	107.5
33	ΔASA	70.9	114.3	69.6	80.5	148.4	44	107.9	142.7	87.5
34	DGh	-0.54	-1.64	-2.97	-3.71	-1.06	-0.59	-3.38	0.32	-2.19
35	G_{hD}	-0.58	-1.91	-6.1	7.37	-1.35	-0.82	-5.57	0.4	-5.97
36	G_{hN}	-0.06	-0.27	-3.11	-3.62	-0.28	-0.23	-2.18	0.07	-1.7
37	ΔH_h	-2.24	-3.43	-4.54	-5.63	-5.11	-1.46	-6.83	-3.84	-5.02
38	$-T\Delta S_h$	1.7	1.79	1.57	1.92	4.05	0.87	3.45	4.16	2.83
39	ΔC_{ph}	14.22	9.41	2.73	3.17	39.06	4.88	20.05	41.98	17.68
40	ΔG_c	0.51	2.71	2.89	3.58	3.22	0.68	3.95	-0.4	1.87
41	ΔH_c	2.77	8.64	4.72	5.69	11.93	1.23	7.64	4.03	3.57
42	$-T\Delta S_c$	-2.25	-5.92	-1.83	-2.11	-8.71	-0.55	-3.69	-4.42	-1.7
43	ΔG	-0.02	1.08	-0.08	-0.13	2.16	0.09	0.56	-0.08	-0.32

44	ΔH	0.51	5.21	0.18	0.05	6.82	-0.23	0.79	0.19	-1.45
45	$-T\Delta S$	-0.54	-4.14	-0.26	-0.19	-4.66	0.31	-0.23	-0.27	1.13
46	V	1	2	4	5	7	0	6	4	5
47	s	0	0	2	3	2	0	2	1	0
48	f	0	1	2	3	2	0	2	2	4

L	M	N	P	Q	R	S	T	V	W	Y
-31.78	-31.18	-30.9	-23.25	-32.6	-26.62	-29.88	-31.23	-30.62	-30.24	-35.01
2.17	1.67	0.09	2.77	0	0.85	0.07	0.07	1.87	3.77	2.67
14.19	13.62	11.72	11.06	11.78	12.4	11.68	12.12	14.73	13.96	13.57
0.13	1.43	3.38	1.58	3.53	52	1.67	1.66	0.13	2.1	1.61
5.98	5.74	5.41	6.3	5.65	10.76	5.68	5.66	5.96	5.89	5.66
2.36	2.28	2.02	1.99	2.17	1.81	2.21	2.1	2.32	2.38	2.2
131	149	132	115	146	174	105	119	117	204	181
21.4	16.25	12.82	17.43	14.45	14.28	9.47	15.77	21.57	21.61	18.03
17.6	14.7	5.4	14.8	9	4.6	6.9	9.5	14.3	17	15
18.78	21.64	13.28	10.93	17.56	26.66	6.35	11.01	13.92	42.53	31.55
1.06	0.64	-0.78	0.12	-0.85	-2.53	-0.18	-0.05	1.08	0.81	0.26
1.32	1.3	1.18	1.24	1.12	0.92	1.3	1.25	1.25	1.03	1.03
0.65	0.65	0.38	0.46	0.4	0.55	0.45	0.52	0.73	0.83	0.65
1.97	1.96	1.56	1.7	1.52	1.48	1.75	1.77	1.98	1.87	1.69
1.21	1.45	0.67	0.57	1.11	0.98	0.77	0.83	1.06	1.08	0.69
1.3	1.05	0.89	0.55	1.1	0.93	0.75	1.19	1.7	1.37	1.47
0.59	0.6	1.56	1.52	0.98	0.95	1.43	0.96	0.5	0.96	1.14
0.69	0.59	1.37	1.61	0.87	1.07	1.34	1.08	0.63	0.76	1.07
35	43	28	22	36	55	20	28	33	61	46
0.79	0.78	1.04	1.16	1.07	1.05	1.13	0.96	0.79	0.77	1.01
0.5	0.42	0.15	0.18	0.11	0.07	0.23	0.23	0.48	0.4	0.26
5.88	5.21	2.12	2.12	2.7	2.53	2.43	2.6	7.14	6.25	3.03
7.37	6.39	5.04	5.65	5.45	5.7	5.53	5.81	7.62	6.98	6.73
1.91	1.25	0.53	0	0.98	0.67	0.7	0.75	1.42	1.33	0.58
0.58	0.73	0.93	2.19	1.2	0.39	0.81	1.25	0.63	1.4	0.72
1.05	1.47	0.93	0	1.63	1.59	0.87	0.46	1.2	0.46	0.52
107.75	105.35	78.01	82.83	93.9	127.34	60.62	76.83	90.78	143.91	123.6
2.19	2.27	1.84	1.32	2.03	1.94	1.57	1.57	1.63	1.9	1.67
4.59	4.14	3.64	3.57	3.06	3.78	3.75	4.09	5.43	4.83	4.93
14.13	13.86	13.02	12.35	12.61	13.1	13.39	12.7	14.56	15.48	13.88
161.4	189.8	134.9	135.1	164.9	210.2	111.4	130.4	143.9	208.8	196.4
31.1	41.3	60.5	60.7	71.5	94.5	48.7	52	28.1	39.5	50.4
129.8	147.9	74	73.5	93.3	116	62.8	78	115.6	167.8	145.9
0.27	-0.6	-3.55	0.32	-3.92	-5.96	-3.82	-1.97	0.13	-3.8	-5.64
0.35	-0.71	-6.63	0.56	-7.12	-12.78	-6.18	-3.66	0.18	-4.71	-8.45
0.07	-0.1	-3.03	0.23	-3.15	-6.85	-2.36	-1.69	0.04	-0.88	-2.82
-3.52	-4.16	-5.68	-1.95	-6.23	-10.43	-5.94	-4.39	-3.15	-8.99	-10.67
3.79	3.56	2.13	2.27	2.31	4.47	2.12	2.42	3.28	5.19	5.03
38.26	31.67	3.91	23.69	3.74	16.66	6.14	16.11	32.58	37.69	30.54
-0.35	1.13	3.26	-0.39	3.69	5.25	3.42	1.74	-0.19	5.59	6.56
3.69	7.06	3.64	1.97	4.47	6.03	5.8	4.42	3.45	13.46	14.41
-4.04	-5.93	-0.39	-2.36	-0.78	-0.78	-2.38	-2.68	-3.64	-7.87	-7.95
-0.08	0.53	-0.3	-0.06	-0.23	-0.71	-0.4	-0.24	-0.06	1.78	0.91
0.17	2.89	-2.03	0.02	-1.76	-4.4	-0.16	0.04	0.3	4.47	3.73
-0.24	-2.36	1.74	-0.08	1.53	3.69	-0.24	-0.28	-0.36	-2.69	-2.82
4	4	4	3	5	7	2	3	3	10	8
2	0	2	0	3	5	0	1	1	2	2
2	3	2	0	3	5	1	1	1	2	2

[a] K^0, compressibility; H_t, thermodynamic transfer hydrophobicity; H_p, surrounding hydrophobicity; *P,* polarity; pH_i , isoelectric point; *pK',* equilibrium constant with reference to the ionization property of COOH group; Mw, molecular weight; B_l, bulkiness; R_f, chromatographic index; *m*, refractive index; H_{nc}, normalized consensus hydrophobicity; E_{sm}, short- and medium-range nonbonded energy; E_l long-range nonbonded energy; E_t,total nonbonded energy (E_{sm} + E_l); P_a, P_b, P_t, and P_c are, respectively, a-helical, b-structure, turn, and coil tendencies; C_a, helical contact area; *F,* mean r.m.s. fluctuational displacement; *Br*, buriedness; R_a, solvent-accessible reduction ratio; *Ns,* average number of surrounding residues; a_n, a_c, and a_m are, respectively, power to be at the N-terminal, C-terminal, and middle of a-helix; *V*°, partial specific volume; N_m and N_1 are, respectively, average medium- and long-range contacts; H_{gm}, combined surrounding hydrophobicity (globular and membrane); ASA_D, ASA_N, and ΔASA are, respectively, solvent-accessible surface area for denatured, native, and unfolding; ΔG_h,\ G_{hD}, and G_{hN} are, respectively, Gibbs free energy change of hydration for unfolding, denatured, and native protein; DH_h, unfolding enthalpy change of hydration; $-T\Delta S_h$, unfolding entropy change of hydration; DC_{ph}, unfolding hydration heat capacity change; ΔG_C, ΔH_C, and $-T\Delta S_C$ are, respectively, unfolding Gibbs free energy, unfolding enthalpy, and unfolding entropy changes of side-chain; ΔG, ΔH, and $-T\Delta S$ are respectively, unfolding Gibbs free energy change, unfolding enthalpy change, and unfolding entropy change of protein; *V*, volume (number of nonhydrogen side-chain atoms); *s,* shape (position of branch point in a side chain); *f*, flexibility (number of side-chain dihedral angles).

[b] K^0 in m^3/mol/Pa (× 10^{-15}); H_t, *Hp*, H_{nc}, H_{gm}, DG_h, G_{hD}, G_{hN}, ΔH_h, $-T\Delta S_h$, ΔG_C, ΔH_c, $-T\Delta S_c$, ΔG, ΔH, and $-T\Delta S$ in kcal/mol; *P* in Debye; P_{hi} and *pK'* in pH units; E_{sm}, E_l, and E_t, in kcal/mol/atom; B_l, C_a, ASA_D, ASA_N, and ΔASA in Å^2; *F* in Å; *V*° in m3/mol (× 10^{-6}); ΔC_{ph} in cal/mol/K; and the rest are dimensionless quantities.

Table 2.2 Appendix. Name and values of the amino acids/residues properties[a,b] used to model kinase and protease inhibition.

Residue	H_t	pK'	R_a	V^0	ASA_N	s	f
A	0.87	2.34	3.7	60.46	33.2	0	0
C	1.52	1.65	3.03	67.7	17.9	0	1
D	0.66	2.01	2.6	73.83	62.4	2	2
E	0.67	2.19	3.3	85.88	81	3	3
F	2.87	1.89	6.6	121.48	33.1	2	2
G	0.1	2.34	3.13	43.25	29.2	0	0
H	0.87	1.82	3.57	98.79	57.7	2	2
I	3.15	1.36	7.69	107.72	28.3	1	2
K	1.64	2.18	1.79	108.5	107.5	0	4
L	2.17	2.36	5.88	107.75	31.1	2	2
M	1.67	2.28	5.21	105.35	41.3	0	3
N	0.09	2.02	2.12	78.01	60.5	2	2
P	2.77	1.99	2.12	82.83	60.7	0	0
Q	0	2.17	2.7	93.9	71.5	3	3
R	0.85	1.81	2.53	127.34	94.5	5	5
S	0.07	2.21	2.43	60.62	48.7	0	1
T	0.07	2.1	2.6	76.83	52	1	1
V	1.87	2.32	7.14	90.78	28.1	1	1
W	3.77	2.38	6.25	143.91	39.5	2	2
Y	2.67	2.2	3.03	123.6	50.4	2	2

a. H_t, thermodynamic transfer hydrophobicity; *pK',* equilibrium constant with reference to the ionization, property of COOH group; R_a, solvent-accessible reduction ratio; $V°$, partial specific volume; ASA_N, solvent-accessible surface area for native protein; *s,* shape (position of branch point in a side chain); *f*, flexibility (number of side-chain dihedral angles).

b. H_t in kcal/mol; *pK'* in pH units; ASA_N in Å^2; $V°$ in m3/mol (× 10^{-6}) and the rest are dimensionless quantities.

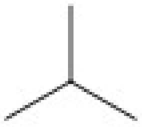
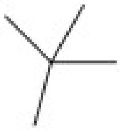
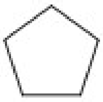
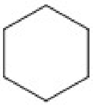
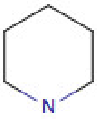
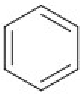
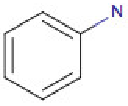
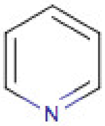
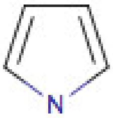
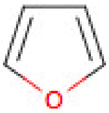
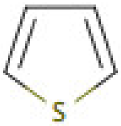
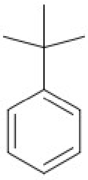
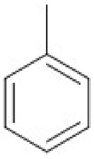
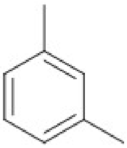
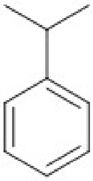
Table 3.1 Appendix. Names of *AASA* vectors in optimum GA-SVM models of conformational stability.

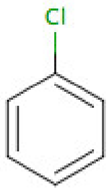
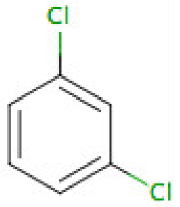
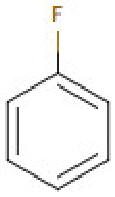
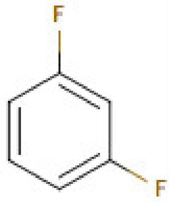
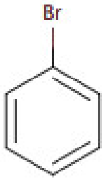
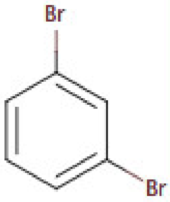
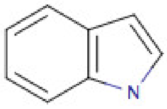
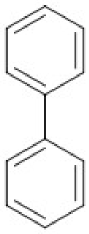
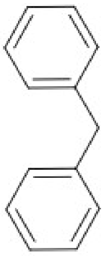
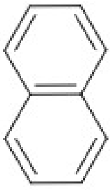
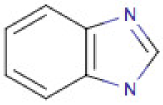
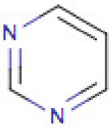
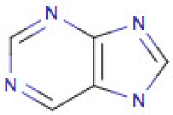
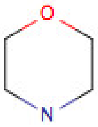
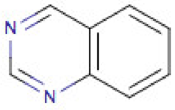
$AASA11_{Nm}$	the amino acids sequence autocorrelation vector at lag 11 weighted by average medium-range contacts;
$AASA8_{P}$	the amino acids sequence autocorrelation vector at lag 8 weighted by polarity;
$AASA7P_{B}$	the amino acids sequence autocorrelation vector at lag 7 weighted by β-structure tendency;
$AASA7G_{hN}$	the amino acids sequence autocorrelation vector at lag 7 weighted by Gibbs Free energy change of hydration for native protein;
$AASA10H_{t}$	the amino acids sequence autocorrelation vector at lag 10 weighted by thermodynamic transfer hydrophobicity;
$AASA14f$	the amino acids sequence autocorrelation vector at lag 14 weighted by flexibility;
$AASA12\Delta G_{C}$	the amino acids sequence autocorrelation vector at lag 12 weighted by unfolding Gibbs Free energy change of side-chain;
$AASA15ASAN$	the amino acids sequence autocorrelation vector at lag 15 weighted by solvent-accessible surface area for native protein;
$AASA15\Delta Cp_{h}$	the amino acids sequence autocorrelation vector at lag 15 weighted by hydration heat capacity change
$AASA14\Delta Cp_{h}$	the amino acids sequence autocorrelation vector at lag 15 weighted by hydration heat capacity

Table 3.2 Appendix. Percent of crossvalidation correct classifications of the optimum SVM model for the *DDG* signs upon mutations according to the mutations type. In brackets the relative fraction of each mutation type in the training dataset of 1383 single point mutants.

Native \ New	Charged	Polar	Apolar
Charged	66% (6%)	79% (9%)	72% (10%)
Polar	70% (5%)	84% (9%)	72% (16%)
Apolar	71% (4%)	88% (13%)	80% (28%)

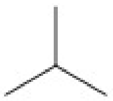
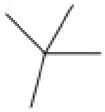
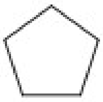
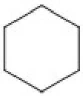
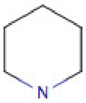
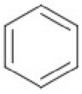
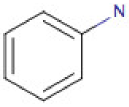
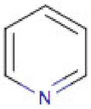
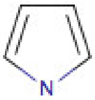
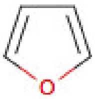
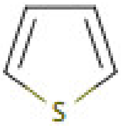
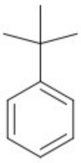
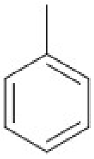
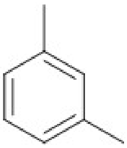
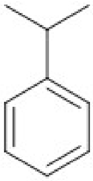
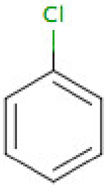
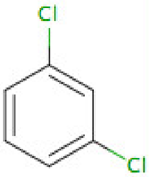
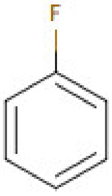
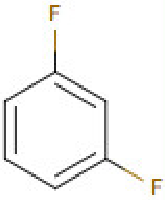
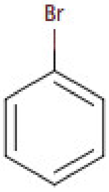
Table 4.1 Appendix. Ligand-wise TFO crossvalidation and prediction accuracies for training and test sets for 30 substructures in the kinase inhibitors dataset according to the optimum topological classifier.

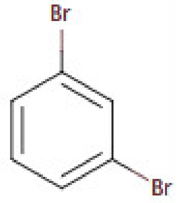
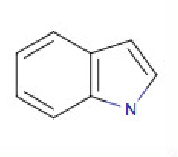
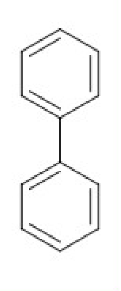
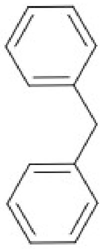
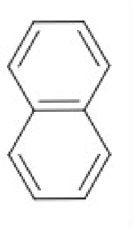
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.83	0.81	1.00	1.00	0.81	0.94	0.80	0.90	0.84	0.86
Q(+)	0.83	0.85	1.00	1.00	0.77	0.92	0.80		0.75	1.00
Q(-)	0.83	0.79			0.86	1.00	0.79	0.90	0.85	0.83
Occ.(%)	14.0	13.2	0.4	0.7	1.3	1.9	1.6	1.1	2.3	3.1
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.81	0.80	0.84	0.85	0.73	0.72	0.79	0.74	0.89	1.00
Q(+)	0.84	0.87	0.95	0.91	0.92	0.92	0.80	0.82	1.00	1.00
Q(-)	0.79	0.76	0.81	0.80	0.50	0.50	0.79	0.70	0.83	1.00
Occ.(%)	91.8	92.5	3.1	2.9	26.1	25.3	19.3	22.8	0.7	1.3
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.86	0.75	0.94	1.00	0.79	0.80	0.81	0.82	0.92	1.00
Q(+)	0.74	0.92	1.00	1.00	0.78	0.85	0.45	0.88	1.00	1.00
Q(-)	0.97	0.60	0.93	1.00	0.80	0.79	0.90	0.80	0.91	1.00
Occ.(%)	2.7	3.1	1.1	0.8	56.8	59.3	3.8	3.7	2.3	2.3

Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.77	0.77	0.69	0.76	0.86	0.84	0.80	0.50	0.86	0.89
Q(+)	0.89	0.94	0.97	0.96	0.90	1.00	1.00	1.00	0.95	0.96
Q(-)	0.66	0.62	0.30	0.40	0.81	0.71	0.67	0.00	0.67	0.65
Occ.(%)	22.3	17.9	7.3	7.9	5.7	4.1	0.2	0.2	10.8	10.9
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2		1.00	0.78	0.75	0.93	1.00	0.74	0.77	0.93	1.00
Q(+)		1.00	0.78	0.78	1.00	1.00	0.68	0.69	1.00	1.00
Q(-)			0.79	0.73	0.92	1.00	0.81	0.89	0.92	1.00
Occ.(%)	0.0	0.1	11.2	13.6	0.93	1.00	2.0	2.4	0.93	1.00
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.96	0.93	0.84	0.83	0.95	0.99	0.89	0.95	0.80	0.83
Q(+)	0.70	0.71	0.91	0.92	0.38	1.00	0.97	1.00	0.91	0.91
Q(-)	0.98	0.97	0.77	0.75	0.98	0.99	0.80	0.83	0.56	0.56
Occ.(%)	5.5	4.9	36.5	36.4	7.8	8.2	2.3	2.1	7.9	8.0

+ and - : the indexes were evaluated for “stable” (IC_{50} < 1 µM) and “unstable” (IC_{50} > 1 µM) kinase inhibition complexes, respectively. Q^2 is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction. Occ.(%) is the occurrence ratio.

Table 5.1 Appendix. Ligand-wise TFO crossvalidation and prediction accuracies for training and test sets for 25 substructures in the protease inhibitors dataset according to the optimum topological classifier.

Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.78	0.78	0.79	0.74	0.90	1.00	0.79	0.75	0.79	0.81
Q(+)	0.77	0.80	0.82	0.85	0.86	1.00	0.75	0.80	0.68	0.83
Q(-)	0.80	0.74	0.71	0.38	1.00		0.86	0.71	0.94	0.78
Occ.(%)	66.8	69.6	8.6	8.0	0.8	2.1	2.7	2.8	5.7	4.9
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.78	0.80	0.91	0.89	0.83	0.78	0.74	0.77	0.75	
Q(+)	0.76	0.81	0.95	0.89	0.79	0.71	0.79	0.86	0.33	
Q(-)	0.81	0.77	0.50		0.87	0.83	0.63	0.43	1.00	
Occ.(%)	96.3	95.6	1.7	2.1	13.0	15.2	9.5	8.2	0.6	0.0
Substr.										
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.53	0.67	1.00	1.00	0.78	0.80	0.73	0.93	0.78	0.89
Q(+)	0.47	1.00	1.00	1.00	0.76	0.82	0.74	0.92	0.84	0.89
Q(-)	0.75	0.50		1.00	0.80	0.77	0.71	1.00	0.58	0.88
Occ.(%)	1.5	0.7	0.8	0.7	93.4	93.9	9.0	9.4	15.6	16.6
Substr.										

	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2	0.83	0.81	0.67	0.75	0.96	0.75	1.00	0.57	0.50	1.00
Q(+)	0.88	0.95	0.88	0.83	0.95	0.79		0.25	0.50	1.00
Q(-)	0.64	0.43	0.43	0.50	1.00	0.67	1.00	1.00		
Occ.(%)	4.9	6.3	1.2	1.9	3.5	4.7	0.3	1.6	0.3	0.9
Substr.	 Br Br		 N							
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Q2	0.67		0.73	0.79	0.75	0.53	0.86	0.82	0.80	0.80
Q(+)	0.50		0.79	0.89	0.77	0.67		0.33	0.78	0.82
Q(-)	1.00		0.60	0.43	0.69	0.00	0.86	1.00	0.83	0.78
Occ.(%)	0.2	0.0	9.3	8.0	4.0	3.5	1.1	2.6	7.2	4.7

+ and - : the indexes were evaluated for “low affinity” (*Ki* > 0.1 µM) and “high affinity” (*Ki* < 0.1 µM) protease inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of high affinity class prediction. Occ.(%) is the occurrence ratio.

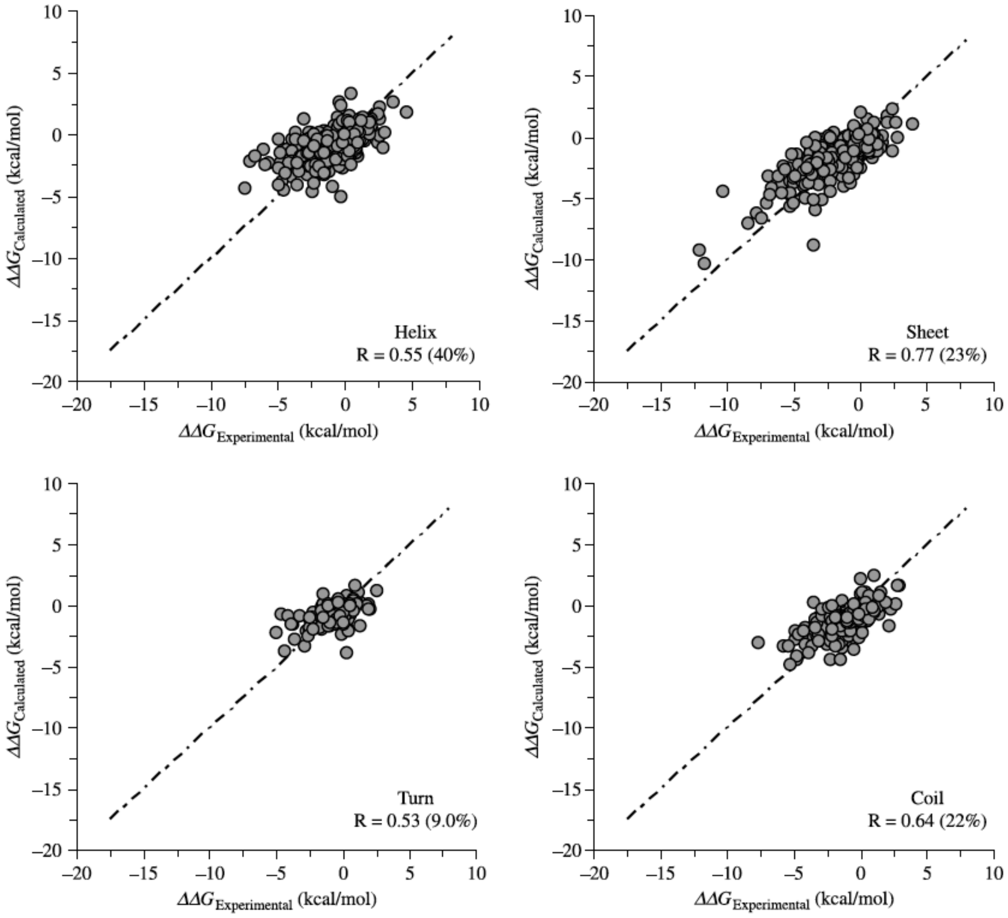


Figure 3.1 Appendix. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change (*ΔΔG*) of protein mutants for each mutation type (secondary structure found in the mutation site) according to regression SVM models including experimental condition data as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

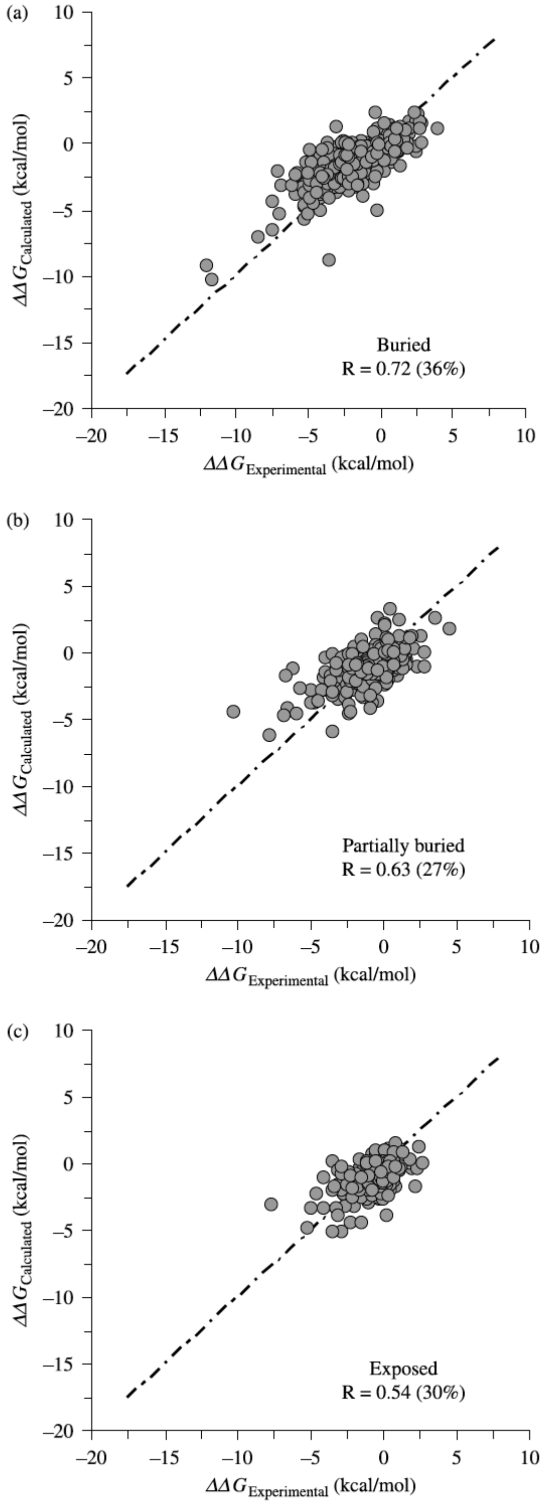


Figure 3.2 Appendix. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change (*ΔΔG*) of protein mutants for each mutation type (accessible surface area (ASA) values of the mutation sites) according to regression SVM models including experimental condition data as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

Publications

1. Fernández M, Fernández L, Sánchez P, Caballero J, Abreu JI: **Proteometric modelling of protein conformational stability using amino acid sequence autocorrelation vectors and genetic algorithm-optimised support vector machines**. *Mol Simulat* 2008, **34**: 857–872.
2. Fernandez M, Shandar A, Sarai A: **Proteochemometric Recognition of Stable Kinase Inhibition Complexes Using Topological Autocorrelation and Support Vector Machines**. *J Chem Inf Model* 2010, **50**:1179–1188.
3. Fernandez M, Caballero J, Fernández L, Sarai A: **Genetic Algorithm-Optimization in Drug Design QSAR: Bayesian Regularized Genetic Neural Networks (BRGNN) and Genetic Algorithm-Optimized Support Vectors Machines (GA-SVM).** *Mol Divers* 2010, doi: 10.1007/s11030-010-9234-9.
4. Fernandez M, Caballero J, Fernández L, Sarai A: **Graphical Representations of Protein Sequences for Alignment-Free Comparative and Predictive Studies. Recognition of Protease Inhibition Pattern from H-Depleted Molecular Graph Representations of Protease Sequences**. *Curr Bioinf* 2010 (*in press*).
5. Fernandez M, Shandar A, Sarai A: **Recognition of high affinity protease inhibition complexes by Autocorrelation Vectors and Support Vector Machines.** *(in preparation)*