5-1-2011

Machine learning and mapping algorithms applied to proteomics problems

William Shane Sanders

MACHINE LEARNING AND MAPPING ALGORITHMS APPLIED TO PROTEOMICS PROBLEMS

By

William Shane Sanders

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Molecular Biology
in the Department of Biochemistry & Molecular Biology

Mississippi State, Mississippi

April 2011

MACHINE LEARNING AND MAPPING ALGORITHMS APPLIED TO

PROTEOMICS PROBLEMS

By

William Shane Sanders

Approved:

Susan M. Bridges Professor Emeritus of Computer Science and Engineering (Director of Dissertation)	Shane C. Burgess Professor of Basic Sciences, College of Veterinary Medicine
Kenneth O. Willeford Professor of Biochemistry and Molecular Biology	John A. Boyle Professor Emeritus of Biochemistry and Molecular Biology
Eric A. Hansen Associate Professor of Computer Science and Engineering	Din-Pow Ma Professor of Biochemistry and Molecular Biology (Graduate Coordinator)
George Hopper Interim Dean of the College of Agriculture and Life Sciences	

Name: William Shane Sanders

Date of Degree: April 29, 2011

Institution: Mississippi State University

Major Field: Molecular Biology

Major Professor: Dr. Susan M. Bridges

Title of Study: MACHINE LEARNING AND MAPPING ALGORITHMS APPLIED TO PROTEOMICS PROBLEMS

Pages in Study: 136

Candidate for Degree of Doctor of Philosophy

Proteins provide evidence that a given gene is expressed, and machine learning algorithms can be applied to various proteomics problems in order to gain information about the underlying biology. This dissertation applies machine learning algorithms to proteomics data in order to predict whether or not a given peptide is observable by mass spectrometry, whether a given peptide can serve as a cell penetrating peptide, and then utilizes the peptides observed through mass spectrometry to aid in the structural annotation of the chicken genome. Peptides observed by mass spectrometry are used to identify proteins, and being able to accurately predict which peptides will be seen can allow researchers to analyze to what extent a given protein is observable. Cell penetrating peptides can possibly be utilized to allow targeted small molecule delivery across cellular membranes and possibly serve a role as drug delivery peptides. Peptides and proteins identified through mass spectrometry can help refine computational gene models and improve structural genome annotations.

DEDICATION

I would like to dedicate this research to my grandmother, Nellie L. Gentry, and to my parents, Patti A. Gentry and William E. Sanders.

ACKNOWLEDGEMENTS

The author expresses his sincere gratitude to the many people without whose selfless assistance this dissertation could not have been produced. Firstly, sincere thanks are due to Dr. Susan M. Bridges, my committee chair, for taking the time and effort to mentor and assist me throughout the doctoral program and dissertation process. The author would also like to express my appreciation for the other members of my dissertation committee, namely Dr. Kenneth O. Willeford, Dr. Shane C. Burgess, Dr. Eric A. Hansen, and Dr. John A. Boyle for their aid and direction throughout the doctoral program. Finally, the author would like to thank the other collaborators and that have assisted with aspects of this research, namely Dr. Fiona M. McCarthy, Dr. Bindu Nanduri, Dr. Nan Wang at the University of Southern Mississippi, Mr. Brandon M. Malone, and Mr. Ken Pendarvis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

With the availability and advancement of rapid genome sequencing technology, an abundance of genomic sequence information is becoming available. In addition, high-throughput proteomics techniques rapidly generate large volumes of data useful for both protein identification and determining protein expression. Since the amount of biological data available for research in the biological sciences is growing rapidly, new experimental and computational methods and tools must be developed to transform data into information. Proteomics focuses on the study of proteins and peptides and the patterns of their expression and regulation within a given organism. Proteins serve as the building blocks of life, and can serve as both structural entities and biochemical catalysts within living cells. Given the rapidly increasing volume of proteomics data, computational techniques for mining and analyzing the data are needed. Machine learning, a subfield within artificial intelligence, is routinely used in computational biology to derive models from large data sets and use these models to predict behavior of an experimental system.

This dissertation applies machine learning and other computational techniques 1) to predict the detectability of peptides using mass spectrometry, 2) to predict the cell

penetration potential of peptides, and 3) to assist in the structural genome annotation of the genomes through a process termed proteogenomic mapping [1].

The first problem, the prediction of peptides detectable by mass spectrometry using machine learning algorithms, specifically examines the use of neural networks in the prediction of detectablility. Published datasets from chicken bursa and lymphoma proteomics experiments were used as training and test sets for the machine learning classifiers. This work has been published in *BMC Bioinformatics*[2].

The second problem, the prediction of peptides capable of penetrating cellular membranes using machine learning algorithms, adapts the features from the prediction of MS peptide dectectability and utilizes these properties in conjunction with support vector machines to predict cell penetration potential. A literature search of known cell-penetrating peptides, along with known cell-penetrating peptides available from commercial vendors was used to create data sets for training and testing the support vector machines. A subset of peptides predicted to be cell-penetrating and non-penetrating were synthesized and utilized for experimental validation of the classifier using avian eukaryotic tissue culture systems in conjunction with fluorescence microscopy and fluorescent quantitative uptake analysis. This work has been submitted for publication in *PLOS Computational Biology* and the manuscript is in revision.

The third problem, using peptides observed by mass spectrometry to assist in the structural annotation of genomes through proteogenomic mapping, has been investigated using the *Gallus gallus* genome. Proteogenomic mapping uses peptides detected from high-throughput mass spectrometry to compliment traditional genome annotation

methods based on computational gene prediction and EST/cDNA libraries. Mapping the peptides to the genome, provides evidence for new functional genomic units that traditional methods often fail to identify. Recent findings from the ENCyclopedia Of DNA Elements (ENCODE) project [3] show that the human genome is more active than previously believed, with a significant portions of the genome being pervasively transcribed. Given this pervasive transcription, it is likely that some of these transcripts are translated into protein. Proteogenomic mapping can reveal which of these transcripts are expressed at the protein level. A paper describing the proteogenomic pipeline has been accepted for publication in *BMC Bioinformatics*. A paper describing the results of proteogenomic mapping with chicken serum is in preparation.

The remainder of this chapter briefly introduces the three problems in more detail and provides an overview of the relevant literature. More in-depth discussions of the relevant literature are included in subsequent chapters along with the research approaches and methodologies, and results.

Prediction of Peptide Properties

The primary amino acid sequences of peptides have been used to calculate and infer a number of properties of peptides such as mass, isoelectric point, secondary structure , etc.. These properties can, in turn, be used by machine learning algorithms to construct classifiers to predict additional peptide properties such as the observability of a given peptide using mass spectrometry or the cell penetrating potential of a given peptide.

Peptides Observability by Mass Spectrometry

In high-throughput non-electrophoretic proteomics, complex mixtures of proteins are subjected to proteolytic digestion with an enzyme such as trypsin before the fragments are separated by liquid chromatography (LC) and analyzed by tandem mass spectrometry. However, for a particular protein, only a portion of the peptides are actually observed experimentally and the set of peptides that are observed from a single protein can vary substantially from one experiment to another. A number of factors contribute to lack of detection of some peptides and to variations in the peptides detected from one experiment to another. These include incomplete proteolytic digestion, small size, poor binding or elution from the type of LC column used, mass range limitations of the mass spectrometer, bias for detecting peptides with an intense MS signal in mixtures, the phenomenon of "ion suppression", the charge prior to ionization, and non-covalent interactions between peptides in the gas phase while in the mass spectrometer [4]. There are also substantial differences in the peptides observed due to experimental variations in protein extraction and/or solublization methods, tissue types, prefractionation, LC separation conditions, and differences between gradients even when the same LC separation conditions are used. Furthermore, different databases, different search software and even different versions of the same software also influence which peptides are detected.

We refer to peptides that can be detected as "flyable". The fact that most proteins in a complex mixture are represented by only a small number of proteolytic peptides presents several difficulties for proteomics researchers [5]. These problems include

assessment of the level of confidence in protein identifications [6], determining the peptide coverage of proteins [7], determining if "missing" proteins are potentially observable [8, 9], and using peptide observability as an adjustment factor for protein quantification based on observed peptides [7, 10]. Recently reported methods for predicting peptide observability have been based on large training datasets from multiple experiments dealing with a single organism [7, 10]. However, because the observability of peptides depends not only on the properties of the peptides themselves but also on specific experimental, instrumental, and analytical procedures, we contend that it is necessary to provide a method for predicting peptide observability for a specific experimental set at the local level. This ability to construct a classifier for a particular dataset is particularly important for researchers who work in smaller laboratories, deal with a variety of organisms and/or tissues, employ a variety of protein extraction protocols, and/or who use a centralized facility for proteomics where they have little control over instrumental and analytical protocols.

We describe a method for constructing a classifier for a proteomics data set that can predict peptide observability for a particular set of experimental conditions. We demonstrate that the classifiers constructed using this method provide critical information for assessing the validity of protein identifications and valuable evidence to support competing hypotheses about the presence or absence of "missing" proteins in a pathway of interest.

The set of tryptic peptides that are observed under experimental conditions can be divided into two classes – proteotrypic and flyable. Proteotrypic peptides are those

experimentally observable peptides that can be used to uniquely identify a protein, while flyable peptides are all peptides that are experimentally observable but may not be proteotrypic [11]. Proteotypic peptides are a subset of flyable peptides and flyable peptides are a subset of all possible tryptic peptides. The spectra generated by mass spectrometry analysis of a complex peptide mixture are matched against theoretical spectra generated from an *in silico* trypsin-digested protein database. The resulting set of peptide identifications is then used for protein identification. Detection of at least one proteotypic peptide is required for protein identification.

There is, however, disagreement among researchers about the number of peptide matches and the peptide coverage of the protein that are required for a protein identification to be considered valid. Protein identifications based on a single proteotypic peptide (sometimes called "one hit wonders") are often viewed with skepticism. Some researchers contend that a protein identification needs at least two proteotypic peptides to be valid, while others contend that a single high quality peptide can be used for identification purposes [6]. Furthermore, some proteins produce only one proteotypic peptide. In addition to the number of peptides identified, the degree of coverage of the protein by peptides may also be used as a measure to assess the validity of the identification—this is typically provided in terms of the percentage of amino acids in the protein "covered" by identified peptides. However, an additional and more meaningful statistic is the percentage of potentially detectable peptides that are observed. This information has the potential to increase (or decrease) the credibility of some single

proteotypic peptides for identification and can prevent loss of important data [6] or the inclusion of erroneous identifications.

Two other research groups have described methods for the prediction of peptide detection using mass spectrometry, but their methods are distinct from ours. Mallick et al. [7] have compiled a large training set from multiple yeast proteomics experiments and built Gaussian mixture discriminant function predictors for a number of different proteomics platforms. Their goal is to characterize the general properties of peptides that can be detected using different proteomics technologies, to determine the coverage of the predicted proteome that is detectable using different technologies, and they also argue that their method can be used to improve protein quantification. Lu et al. [10] describe a classifier for predicting peptide observability that is a component of a method for absolute protein quantification and that adjusts scores for protein abundance based on the predicted detectability of *in silico* generated tryptic peptides.

Cell Penetrating Peptides

Cell penetrating peptides (CPPs), also referred to as "Trojan" peptides, protein transduction domains, or membrane translocation sequences, are typically hydrophobic linear arrangements of 8-24 amino acids able to cross the lipid bi-layer membrane that serves as the cell's outer barrier and gain access to the interior of the cell and its components [12]. Penetratin, an Antennapedia derived peptide, and the HIV derived Tat peptide were some of the first commonly studied CPPs, and along with transportan peptides (derived from galanin receptor ligand proteins), make up three major families of CPPs. The remainder of CPPs are classified in a fourth, miscellaneous family [12].

Cell penetrating peptides capable of transporting other active molecules inside the cell have the potential to serve as drug delivery peptides. Although there is some controversy regarding CPPs as drug delivery systems because of their lack of specificity for cell type, the general consensus among researchers is that both general CPPs and cell-specific CPPs will be developed into effective drug delivery systems in the future [13, 14]. A classification system that can determine whether or not a peptide can serve as a CPP can enable researchers to quickly screen candidate molecules for their potential viability for use in a customizable drug delivery regime.

Much of the previous work in the prediction of CPPs has involved the use of a set of composite features assembled from primary biochemical properties through the use of principal component analysis [15-17]. These composite features, or *z*-scores, consist of a numerical value and an associated range. To predict cell-penetrating capability of a candidate peptide, the *z*-scores are computed for the peptide, and, if the *z*-scores fall within the range of known CPP *z*-scores, the peptide is classified as cell-penetrating [16, 17]. While this method has a high accuracy (>95% correct prediction of novel CPPs) for generating novel CPPs [16], it performs rather poorly (68% correct prediction) when trying to distinguish known non-penetrating peptides that are closely related to known CPPs [17] and yields little information about exactly which biochemical properties contribute to the difference between these two classes. More recent work examines the use of quantitative structure-activity relationship (QSAR) derived features to predict penetration potential. The training process iteratively removes sequences that are difficult to classify and thus the classification accuracies reported are biased [18].

Further research into this topic is necessary to allow potential drug delivery peptides to be rapidly screened for usefulness.

Using the basic biochemical properties of peptides as features instead of composite *z*-scores can potentially provide more insight into the differences between the class of CPPs and non-penetrating peptides when coupled with the use of a machine learning classifier such as a support vector machine. Additionally, once trained, these machine learning classifiers can then be used for rapid screening of candidate CPPs prior to their synthesis.

Proteogenomic Mapping

Structural genome annotation is the process of identifying all of the structural elements that comprise an organism's sequenced genome. These structural elements can include regions that code for proteins, both coding and non-coding RNAs, regulatory regions, and DNA binding motifs. Traditionally, this has been accomplished through the use of expressed sequence tags (ESTs) and cDNA libraries, transcribed RNA that is reverse translated into DNA sequences. These ESTs and cDNAs generally represent approximately 500-800 base pair mRNA sequences that are sequenced as they are, or translated back into cDNA and then sequenced [19, 20]. These EST and cDNA libraries are then aligned with the sequenced genome to identify regions representing exons and whole genes that are actively transcribed [19, 20].

These library based-methods are typically complemented by the use of computational gene finders. The computational gene finders use the EST and cDNA libraries to identify patterns within the genome indicative of coding regions. This is

known as homology based computational annotation [19, 20]. Some programs can also perform *de novo* based genome annotation where they detect signal information within the genome and use these signals to predict coding regions [19]. Computational gene prediction tools are known to produce a number of errors and significant resources are dedicated to identifying and correcting these errors in genome annotation projects [19, 21]. It has been estimated that the exact genomic structure is only correctly identified by computational gene finders 50-60% of the time within the human genome, the most well sequenced and annotated genome [21]. Both homology-based methods and *de novo* methods are effective for identifying new genetic sequences similar to known genes or with known signals. However, these methods are ineffective for identifying new genes with limited sequence similarity or signal information [19]. The use of high throughput proteomics, in conjunction with the genome sequence, has the potential to provide additional evidence for new genes or corrections to the boundaries of known genes.

The use of high throughput shotgun proteomics data derived from mass spectrometry experiments is increasingly being used as a complementary method for structural genome annotation [22]. This use of proteomics data to aid in genome annotation began around 2001[23] for several prokaryotic projects, and was popularized in 2004 by Jaffe et al., who coined the term proteogenomic mapping [1]. Proteomic evidence, identified as expressed Protein Sequence Tags (ePSTs), provides proof that a given gene is expressed, and when back translated and aligned with the sequenced genome, provide structural annotation information for a genome's functional elements [22, 24]. This can include *"confirmation of translation, reading-frame determination,*

identification of gene and exon boundaries, evidence for post translational processing, identification of splice-forms including alternative splicing, and also, the prediction of completely novel genes" [22]. Since the development of proteogenomic mapping, it has been utilized in a number of both prokaryotic [1, 23, 25-31] and eukaryotic [9, 32-42] genome annotation projects, and is increasingly becoming a part of standard annotation pipelines utilizing multiple sources of evidence (sequenced nucleic acids, computational gene prediction, and proteomics data) [20].

Summary

This dissertation uses proteomics data combined with machine learning tools to contribute to the prediction of peptide properties and to improve the structural annotation of the chicken genome. The dissertation demonstrates the use of proteomics data and machine learning to solve three different bioinformatics problems. The remainder of this dissertation reviews the relevant literature of the three proteomics problems, describes and discusses the research performed, and presents the results of that research.

LITERATURE CITED

1. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation**. *Proteomics* 2004, **4**(1):59-77.

2. Sanders WS, Bridges SM, McCarthy F, Nanduri B, Burgess SC: **Prediction of peptides observable by mass spectrometry applied at the experimental level.** *BMC Bioinformatics* 2007:In press.

3. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.

4. Hernandez H, Robinson CV: **Dynamic protein complexes: insights from mass spectrometry**. *J Biol Chem* 2001, **276**(50):46685-46688.

5. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.

6. Veenstra TD, Conrads TP, Issaq HJ: **What to do with "one-hit wonders"?** *Electrophoresis* 2004, **25**(9):1278-1279.

7. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T *et al*: **Computational prediction of proteotypic peptides for quantitative proteomics**. *Nat Biotechnol* 2007, **25**(1):125-131.

8. Buza JJ, Burgess SC: **Modeling the proteome of a Marek's disease transformed cell line: a natural animal model for CD30 overexpressing lymphomas**. *Proteomics* 2007, **7**(8):1316-1326.

9. McCarthy FM, Cooksey AM, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a whole organ using proteomics: the avian bursa of Fabricius**. *Proteomics* 2006, **6**(9):2759-2771.

10. Lu P, Vogel C, Wang R, Yao X, Marcotte EM: **Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation**. *Nat Biotechnol* 2007, **25**(1):117-124.

11. Kuster B, Schirle M, Mallick P, Aebersold R: **Scoring proteomes with proteotypic peptide probes**. *Nat Rev Mol Cell Biol* 2005, **6**(7):577-583.

12. Kilk K: **Cell-penetrating peptides and bioactive cargoes. Strategies and mechanisms.** Stockholm: Stockholm University; 2004.

13. Schwartz JJ, Zhang S: **Peptide-mediated cellular delivery**. *Curr Opin Mol Ther* 2000, **2**(2):162-167.

14. Vives E: **Present and future of cell-penetrating peptide mediated delivery systems: "is the Trojan horse too wild to go only to Troy?"**. *J Control Release* 2005, **109**(1-3):77-85.

15. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids**. *J Med Chem* 1998, **41**(14):2481-2491.

16. Hallbrink M, Kilk K, Elmquist A, Lundberg P, Lindgren M, Jiang Y, Pooga M, Soomets U, Langel U: **Prediction of Cell-Penetrating Peptides**. *International Journal of Peptide Research and Therapeutics* 2005, **11**(4):249-259.

17. Hansen M, Kilk K, Langel U: **Predicting cell-penetrating peptides**. *Adv Drug Deliv Rev* 2008, **60**(4-5):572-579.

18. Dobchev DA, Mager I, Tulp I, Karelson G, Tamm T, Tamm K, Janes J, Langel U, Karelson M: **Prediction of Cell-Penetrating Peptides Using Artificial Neural Networks**. *Curr Comput Aided Drug Des*, **2010**:6.

19. Mathe C, Sagot MF, Schiex T, Rouze P: **Current methods of gene prediction, their strengths and weaknesses**. *Nucleic Acids Res* 2002, **30**(19):4103-4117.

20. Allen JE, Pertea M, Salzberg SL: **Computational gene prediction using multiple sources of evidence**. *Genome Res* 2004, **14**(1):142-148.

21. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L: **Identification and correction of abnormal, incomplete and mispredicted proteins in public databases**. *BMC Bioinformatics* 2008, **9**(353):353.

22. Castellana N, Bafna V: **Proteogenomics to discover the full coding content of genomes: a computational perspective**. *J Proteomics* 2010, **73**(11):2124-2135.

23. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Matching peptide mass spectra to EST and genomic DNA databases**. *Trends Biotechnol* 2001, **19**(10 Suppl):S17-22.

24. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD: **Proteogenomics: needs and roles to be filled by proteomics in genome annotation**. *Brief Funct Genomic Proteomic* 2008, **7**(1):50-62.

25. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N *et al*: **The complete genome and proteome of Mycoplasma mobile**. *Genome Res* 2004, **14**(8):1447-1461.

26. Savidor A, Donahoo RS, Hurtado-Gonzales O, Verberkmoes NC, Shah MB, Lamour KH, McDonald WH: **Expressed peptide tags: an additional layer of data for genome annotation**. *J Proteome Res* 2006, **5**(11):3048-3058.

27. Wilkins MJ, Verberkmoes NC, Williams KH, Callister SJ, Mouser PJ, Elifantz H, N'Guessan A L, Thomas BC, Nicora CD, Shah MB *et al*: **Proteogenomic monitoring of Geobacter physiology during stimulated uranium bioremediation**. *Appl Environ Microbiol* 2009, **75**(20):6591-6599.

28. Kunec D, Nanduri B, Burgess SC: **Experimental annotation of channel catfish virus by probabilistic proteogenomic mapping**. *Proteomics* 2009, **9**(10):2634-2647.

29. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C, Lecompte O: **Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol**. *Genome Res* 2009, **19**(1):128-135.

30. Nanduri B, Wang N, Lawrence ML, Bridges SM, Burgess SC: **Gene model detection using mass spectrometry**. *Methods Mol Biol* 2010, **604**:137-144.

31. Payne SH, Huang ST, Pieper R: **A proteogenomic update to Yersinia: enhancing genome annotation**. *BMC Genomics* 2010, **11**(460):460.

32. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S *et al*: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry**. *Genome Biol* 2005, **6**(1):R9.

33. Matis M, Zakelj-Mavric M, Peter-Katalinic J: **Mass spectrometry and database search in the analysis of proteins from the fungus Pleurotus ostreatus**. *Proteomics* 2005, **5**(1):67-75.

34. Smith JC, Northey JG, Garg J, Pearlman RE, Siu KW: **Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of Tetrahymena thermophila**. *J Proteome Res* 2005, **4**(3):909-919.

35. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ: **Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics**. *Genome Biol* 2006, **7**(4):R35.

36. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics**. *Proc Natl Acad Sci U S A* 2008, **105**(52):21034-21038.

37. Colinge J, Cusin I, Reffas S, Mahe E, Niknejad A, Rey PA, Mattou H, Moniatte M, Bougueleret L: **Experiments in searching small proteins in unannotated large eukaryotic genomes**. *J Proteome Res* 2005, **4**(1):167-174.

38. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry**. *Genome Res* 2007, **17**(2):231-239.

39. Kalume DE, Peri S, Reddy R, Zhong J, Okulate M, Kumar N, Pandey A: **Genome annotation of Anopheles gambiae using mass spectrometry-derived data**. *BMC Genomics* 2005, **6**(128):128.

40. Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, Stephenson JL, Jr.: **Whole genome searching with shotgun proteomic data: applications for genome annotation**. *J Proteome Res* 2008, **7**(1):80-88.

41. Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ: **Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations**. *Genome Res* 2008, **18**(10):1660-1669.

42. Lucitt MB, Price TS, Pizarro A, Wu W, Yocum AK, Seiler C, Pack MA, Blair IA, Fitzgerald GA, Grosser T: **Analysis of the zebrafish proteome during embryonic development**. *Mol Cell Proteomics* 2008, **7**(5):981-994.

CHAPTER II

PREDICTION OF PEPTIDES OBSERVABLE BY MASS SPECTROMETRY APPLIED AT THE EXPERIMENTAL SET LEVEL

Abstract

Background

When proteins are subjected to proteolytic digestion and analyzed by mass spectrometry using a method such as 2D LC MS/MS, only a portion of the proteotypic peptides associated with each protein will be observed. A number of factors can contribute to the inability to detect some peptides including protein extraction methods, choice of proteolytic enzymes, properties of the peptides, experimental and instrumentation conditions, non-covalent interactions by the peptides in the gas phase, and changes to database search algorithms. The ability to predict which peptides can and cannot potentially be observed for a particular experimental dataset has several important applications in proteomics research including calculation of peptide coverage in terms of potentially detectable peptides, systems biology analysis of data sets, and protein quantification.

Results

We have developed a methodology for constructing artificial neural networks that can be used to predict which peptides are potentially observable for a given set of experimental, instrumental, and analytical conditions for 2D LC MS/MS (a.k.a Multidimensional Protein Identification Technology [MudPIT]) datasets. Neural network classifiers constructed using this procedure for two MudPIT datasets exhibit 10-fold cross validation accuracy of about 80%. We show that a classifier constructed for one dataset has poor predictive performance with the other dataset, thus demonstrating the need for dataset specific classifiers. Classification results with each dataset are used to compute informative percent amino acid coverage statistics for each protein in terms of the predicted detectable peptides in addition to the percent coverage of the complete sequence. We also demonstrate the utility of predicted peptide observability for systems analysis to help determine if proteins that were expected but not observed generate sufficient peptides for detection.

Conclusions

Classifiers that accurately predict the likelihood of detecting proteotypic peptides by mass spectrometry provide proteomics researchers with powerful new approaches for data analysis. We demonstrate that the procedure we have developed for building a classifier based on an individual experimental data set results in classifiers with accuracy comparable to those reported in the literature based on large training sets collected from multiple experiments. Our approach allows the researcher to construct a classifier that is

specific for the experimental, instrument, and analytical conditions of a single experiment and amenable to local, condition-specific, implementation. The resulting classifiers have application in a number of areas such as determination of peptide coverage for protein identification, pathway analysis, and protein quantification.

Background

In high-throughput non-electrophoretic proteomics complex mixtures of proteins are subjected to proteolytic digestion with an enzyme such as trypsin before the fragments are separated by liquid chromatography (LC) and analyzed by tandem mass spectrometry. However, for a particular protein, only a portion of the peptides are actually observed experimentally and the set of peptides that are observed from a single protein can vary substantially from one experiment to another. A number of factors contribute to the inability to detect some peptides and to variations in the peptides that are detected from one experiment to another. These include incomplete proteolytic digestion, small size, poor binding or elution from the type of LC column used, the limited mass range that can be detected by the mass spectrometer, bias toward detecting peptides with an intense MS signal in mixtures, the phenomenon of "ion suppression", the charge prior to ionization, and non-covalent interactions between peptides in the gas phase while in the mass spectrometer [1]. In addition, there are substantial differences in the peptides observed due to variations in the protein extraction and or solublization methods, tissue types, prefractionation, LC separation conditions, and differences between gradients even when the same LC separation conditions are used. Furthermore,

different databases, different search software and even different versions of the same software also influence which peptides that are detected.

We refer to peptides that can be detected as "flyable". The fact that most proteins in a complex mixture are represented by only a small number of proteolytic peptides presents several difficulties for proteomics researchers [2]. These problems include assessment of the level of confidence in protein identifications [3], determining the peptide coverage of proteins [4], determining if "missing" proteins are potentially observable [5, 6], and using peptide observability as an adjustment factor for protein quantification based on observed peptides [4, 7]. Recently reported methods for predicting peptide observability have been based on large training datasets from multiple experiments dealing with a single organism [4, 7]. However, because the observability of peptides depends not only on the properties of the peptides themselves but also on specific experimental, instrumental, and analytical procedures, we contend that it is necessary to provide a method for predicting peptide observability for a specific experimental set at the local level. This ability to construct a classifier for a particular dataset is particularly important for researchers who work in smaller laboratories, deal with a variety of organisms and/or tissues, employ a variety of protein extraction protocols, and/or who use a centralized facility for proteomics where they have little control over instrumental and analytical protocols.

Here we describe a method for constructing a classifier for a proteomics data set that can predict peptide observability for a particular set of experimental conditions. We demonstrate that the classifiers constructed using this method provide critical information

for assessing the validity of protein identifications and valuable evidence to support competing hypotheses about the presence or absence of "missing" proteins in a pathway of interest.

The set of tryptic peptides that are observed under experimental conditions can be divided into two classes – proteotrypic and flyable. Proteotrypic peptides are those experimentally observable peptides that can be used to uniquely identify a protein, while flyable peptides are all peptides that are experimentally observable but may not be proteotrypic [8]. Proteotypic peptides are a subset of flyable peptides and flyable peptides are a subset of all possible tryptic peptides. The spectra generated by mass spectrometry analysis of a complex peptide mixture are matched against theoretical spectra generated from an *in silico* trypsin-digested protein database. The resulting set of peptide identifications is then used for protein identification. By definition, detection of at least one proteotypic peptide is required for protein identification.

There is, however, disagreement among researchers about the number of peptide matches and the peptide coverage of the protein that are required for an identification to be considered valid. Protein identifications based on a single proteotypic peptide (sometimes called "one hit wonders") are often viewed with skepticism. Some researchers contend that a protein identification needs at least two proteotypic peptides to be valid, while others contend that a single high quality peptide can be used for identification purposes [3]. Furthermore, some proteins produce only one proteotypic peptide. In addition to the number of peptides identified, the degree of coverage of the protein by peptides may also be used as a measure to assess the validity of the

identification—this is typically provided in terms of the percentage of amino acids in the protein "covered" by identified peptides. However, an additional and more meaningful statistic is the percentage of potentially detectable peptides that are observed. This information has the potential to increase (or decrease) the credibility of some single proteotypic peptides for identification and can prevent loss of important data [3] or the inclusion of erroneous identifications.

Researchers using proteomics are interested in not only cataloging proteins present, but also in studying the location and differential expression of the proteins involved in biochemical pathways [2]. Often, one or more proteins referenced to participate in a canonical pathway are not observed in a proteomics dataset, but most other proteins in the pathway are present [5, 6]. Conversely, a protein that has never been identified in that pathway may be identified by a single proteotypic peptide. In the first case, it is important to know whether these missing proteins generate a sufficient number of potentially observable proteotypic peptides to support identification under the specific experimental conditions or whether the protein truly appears to be absent. In the second case, it is important to determine if a protein may reasonably be expected to be identified by only one peptide under the experimental conditions—an identification of a protein with a single peptide where the protein is predicted to produce many observable proteotypic peptides should be viewed with suspicion.

Two recently published papers describe methods for the prediction of peptide detection using mass spectrometry, but their methods are distinct from ours. Mallick et al. [4] have compiled a large training set from multiple yeast proteomics experiments

and built Gaussian mixture discriminant function predictors for a number of different proteomics platforms. Their goal is to characterize the general properties of peptides that can be detected using different proteomics technologies, to determine the coverage of the predicted proteome that is detectable using different technologies, and they also argue that their method can be used to improve protein quantification. Lu et al. [7] describe a classifier for predicting peptide observability that is a component of a method for absolute protein quantification and that adjusts scores for protein abundance based on the predicted detectability of *in silico* generated tryptic peptides. In contrast, our procedure is specifically developed for generating a classifier for a single data set to predict flyable peptides for a particular set of experimental conditions (biological sample, protein extraction protocol, mass spectrometric instrumentation, HPLC column type, database search algorithm and settings, etc.) and to be applied locally. We demonstrate that the resulting classification provides valuable information with regard to peptide coverage of a protein and can assist the proteomics researcher in a systems analysis of the dataset.

Results and Discussion

We have developed a procedure for building a classifier to predict peptide flyability from a proteomics dataset. The output of the protein identification algorithms for a proteomics dataset includes the proteins that were identified and the peptides that were used for each protein identification. As Figure 2.1 illustrates, the classifier construction process includes selection of a set of observed and unobserved peptides for the training set, extraction of features to represent the peptides in the training set,

normalization of the feature values, feature subset selection, and training and testing of the classifier.

Training Set Compilation Strategy

The first step in the process is selection of a set of peptides for the training data set. The naïve approach is to use all observed peptides for the positive examples and all non-observed *in silico* generated peptides from identified proteins for the negative examples. However, this approach ignores several complications that arise when processing proteomics datasets. First, some of the "observed" peptides will be false positive identifications. The probability that a peptide is a false positive identification is greatly reduced if it is one of multiple peptides used to identify a protein since the probability of this occurring by chance is small [3]. Therefore, we limit the positive examples to the peptides associated with proteins that were identified using multiple unique peptides. Peptides chosen for negative examples are also limited to the set of proteins identified by multiple peptides. However, selection of negative examples is also complicated by the fact that the number peptides observed for a protein is directly related to protein abundance in the sample. Isotope-free quantification methods for proteomics datasets make use of the relationship between the number of peptides observed and protein concentration [9-12]. To avoid the problem of labeling peptides that were not observed as negative examples because they are associated with low abundance proteins, we have chosen to compile the negative examples from the proteins that were identified with the largest number of peptides. Although this introduces a bias for peptides from abundant and large proteins, this strategy insures, to the extent possible, that the peptides used for

negative examples were present in sufficient quantity to be potentially observable. We have developed the following procedure for selection of the training set to ensure that the peptides selected for the class of observable peptides are high confidence identifications and that the peptides selected for the negative examples are truly "unobservable" under the specific experimental conditions.

1. Rank the protein identifications by the number of peptides used in the identification and include only identifications based at least two distinct peptides.

2. Retrieve the amino acid sequence for each of the proteins in step 1, perform *in silico* trypsin (or appropriate enzyme) digestion of the proteins, and compile a list of all predicted tryptic peptides of length greater than 6 amino acids (because this number gives a probability of the sequence identifying another sequence at random of 1 in 19^6 and which is reasonable for a eukaryote genome of around 4 billion base-pairs such as human).

3. If a peptide is present in the experimental data, it is assigned a value of 1 and if it is not observed in the experimental data it is given a value of 0. There will be many more with a value of 0 than with 1.

4. The peptides labeled with a 1 in the previous step are used as the positive examples in the training set. Suppose the size of this set is *n*. In order assure that peptides used as negative examples were present in sufficient quantity for detection and to also help produce a balanced training set, we select the first *n* "unobserved" peptides from the proteins ranked by the number of peptides used for identification.

Feature Generation and Classifier Construction

Our approach for generating features to represent each peptide in the training set uses both the features listed in Table 2.1 (called Feature Set 1) and features constructed using properties from the AAIndex [13]. The first set of features (see Table 2.1) includes basic properties of the peptide (e.g. mass and size) and features related to the amino acid composition of the peptide. The AA Index is a compilation in a set of tables of 544 different indices used to characterize amino acids. It includes indices for wide

variety of characteristics of amino acids including hydrophobicity, participation in certain types of structures, etc. A feature value was generated for each peptide for each index representing the sum of the index values for all amino acids in the peptide. Combination of Feature Set 1 and the AAindex features results in a total of 596 features for each peptide. Although this set includes a large number of redundant features, we have shown that using both sets as input for the feature selection process yields improved classifier performance over use of each feature set alone. For example, with the avian bursal dataset described below, the 10-fold classification accuracy of neural networks built with the AAIndex features only is 72%, with Feature Set 1 only is 71%, and with both feature sets is 81%. Because the values of the features cover a wide range of numeric values, NV normalization is used to make the numeric range of all features 0-1. Feature subset selection is then performed to find the set of feature most relevant to the task of predicting flyability and to remove redundant and non-informative features. We use a feature selection method that performs a greedy search through feature space to identify features based on the level of consistency with class values when the training data is compared to the entire set of attributes [14]. The reduced set of features is used to train the classifier. A 3-layer neural network classifier is constructed with an input unit for each of the selected features, $(i+1)/2$ hidden units where i is the number of input units, and a single output unit. The neural network is trained using the training set constructed with the strategy described above and tested using 10-fold cross validation. Multilayer neural networks provide a robust method for learning a functional mapping from numeric

attribute values to a class value—in this case a mapping from numeric features describing the peptide to the classes "observable" and "unobservable."

In order to demonstrate the utility of our approach, we have used the methodology described above to build classifiers for two different published MudPIT data sets: 1) an avian bursa of Fabricius data set consisting of 5198 proteins [6], and 2) a Hodgkin's lymphoma model data set consisting of 3983 proteins [5]. The classifiers built using our procedure had 10-fold cross validation classification accuracies of 81% and 72% respectively. Table 2.2 lists the features selected that best distinguish observed peptides from unobservable peptides for both datasets. Table 2.3 reports the accuracy and confusion matrices for the neural networks for both data sets based on 10-fold cross validation.

The features selected tend to be related to structural properties of the peptides. For example, consider the features selected for the avian bursa classifier. Prolines tend to break alpha helices and prolines located adjacent to lysine or arginine also interfere with trypsin digestion. Amino acids with small side chains such as glycine and alanine increase the flexibility of the peptide. The charge, polarity, hydrophobicity, and the behavior of the peptide in solvent also influence flyability.

Our classifiers achieved classification accuracies comparable to the rates reported by Mallick et al. [4] and Lu et al. [7] for much simpler yeast systems. The accuracy statistics reported by Mallick et al. are difficult to compare to ours because they report specificity in terms of (1 - positive predictive ratio) where the positive predictive ratio is defined as (true positives/(true positives + false positives)) rather than the more

traditional true positive ratio (true positives/(true positives + false negatives). Lu et al. report a 69% true positive rate for observed and a 90% true positive rate for non-observed. Note that it is possible to achieve an 82% true positive rate for the non-observed class for their classifier by guessing non-observed in every case. In addition, they include very small peptides (3 -5 aa) in their analysis and we exclude peptides of this length from our study because of the high probability of random matches to multiple proteins and their lack of power as unique identifiers.

In order to evaluate the importance of building classifiers that are specific for a particular dataset, we tested each of the classifiers above with the data used for training the other classifier (i.e. avian bursal classifier with Hodgkin's lymphoma model data set as test set and vice versa). The results (Table 2.4) demonstrate that there is a substantial loss of classifier accuracy when using a classifier trained with one data set to predict peptide observability with the other data set. In both cases, the true positive rate (prediction of observability) decreased dramatically (almost to the level that would be achieved by random guessing). These results are consistent with those reported by Mallick et al. [4] when a classifier trained with yeast data was used to predict observability with human data. These results clearly demonstrate the need for classifiers to be trained for each experimental set.

We use the two classifiers described above for the avian bursa dataset and the Hodgkin's lymphoma model dataset to demonstrate the utility of the classifiers for calculating an informative peptide coverage statistic for proteins and for analysis of system's biology datasets. In Table 2.4 the sections in white show, for a subset of

proteins that were observed in the data, the total number of tryptic peptides generated by *in silico* tryptic digestion, the number observed, the number of peptides predicted to be detectable by each classifier, and the amino acid coverage both in terms of the total number of tryptic peptides and in terms of those predicted to be observable. As expected, in most cases the amino acid coverage for peptides predicted to be detectable is higher, sometimes substantially higher, than the total amino acid coverage. In general, this approach allows the researcher to determine how many peptides might reasonably be expected to be detected.

We have also used the bursal neural network and the Hodgkin's lymphoma model neural network to determine if proteins that are "missing" from a pathway of interest are likely to be potentially observable. The results are given in Table 2.5. As McCarthy et al. [6] reported, most components of the programmed cell death pathway with known orthologs in chicken were observed in the avian bursa data set with the exception of the protein DR3. The peptides produced by *in silico* tryptic digestion of DR3 (GI 118106991) were used as input to our neural network for this data set. As shown in Table 2.5 (yellow section), none of the peptides for this protein were predicted to be observable. In contrast, for proteins that were observed, the average number of observable peptides was 5. For the Hodgkin's lymphoma model dataset, there were five proteins that we expected to observe because we have observed them using other methods in other experiments [15, 16] but we did not see them in this experiment (shown in yellow in Table 2.4). The results in Table 2.5 show that none of the tryptic peptides for these proteins is predicted to be observable under the given experimental conditions

while a set of proteins of similar size that were observed were predicted to be observable. Although these results cannot be used to demonstrate conclusively that a protein does or does not exist in a data set, they can be used as one piece of evidence to confirm or refute a hypothesis about the presence of a protein under certain conditions and to plan further wet lab experiments.

Conclusions

We present a procedure for constructing a classifier to predict which tryptic peptides in a protein are likely to be detectable by mass spectrometry for a specific set of experimental and instrumental conditions. We demonstrate that it is possible to construct a classifier with accuracy comparable to those previously reported based on the accumulation of large training sets from multiple experiments. We also show that a classifier constructed based on one dataset does not perform at an acceptable level when predicting observability for another dataset and thus it is necessary to construct classifiers that are specific for one set of experimental conditions. The resulting classifier provides researchers with a tool that can provide information about peptide coverage of proteins in terms of which proteins are likely to be detectable. It can also be used as one line of evidence in a systems analysis to evaluate alternative hypotheses concerning proteins that were not observed but that were expected. If the "missing" protein generates many predicted detectable peptides but none were observed, then this provides additional probabilistic evidence of absence of the protein—a very difficult hypothesis to demonstrate conclusively. The classifier allows researchers to distinguish between proteins that are not likely to be detected with the methodology versus proteins that were

not expressed in the biological system. Only by making this distinction is it possible to accurately interpret proteomics results and improve biological modeling.

Methods

Biological Samples

Methods used to collect the biological samples, analyze the samples using mass spectrometry, and identify proteins are described in detail in[5] and [6]. All samples were analyzed by MudPIT using an LCQ Deca XP Plus IT mass spectrometer and database search was conducted using TurboSEQUEST (Bioworks Browser; ThermoElectron).

Software

Custom Perl scripts were written to extract the accessions of proteins and lists of peptides from Sequest output files, to query NCBI and download the protein sequences, to trypsin digest the proteins, to determine which peptides had been observed in the dataset, to select the positive and negative peptides for the data sets, and to compute the feature vectors for each peptide. The software implements the rules for trypsin digestion described for the ExPASy PeptideCutter tool [17]. WEKA Explorer Version 3.4.10, a software package containing a collection of machine learning algorithms for data mining available at **http://www.cs.waikato.ac.nz/ml/weka/** [14] was used for feature selection, and building and testing the classifier. The software that generates a training set from a Sequest output file and a detailed readme describing how to generate classifiers for a

specific dataset using Weka is available for download in the Tools section of AgBase (www.agbase.msstate.edu).

LITERATURE CITED

1. Hernandez H, Robinson CV: **Dynamic protein complexes: insights from mass spectrometry**. *J Biol Chem* 2001, **276**(50):46685-46688.

2. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.

3. Veenstra TD, Conrads TP, Issaq HJ: **What to do with "one-hit wonders"?** *Electrophoresis* 2004, **25**(9):1278-1279.

4. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T et al: **Computational prediction of proteotypic peptides for quantitative proteomics.** *Nat Biotechnol* 2007, **25**(1):125-131.

5. Buza JJ, Burgess SC: **Modeling the proteome of a Marek's disease transformed cell line: a natural animal model for CD30 overexpressing lymphomas**. *Proteomics* 2007, **7**(8):1316-1326.

6. McCarthy FM, Cooksey AM, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a whole organ using proteomics: the avian bursa of Fabricius.** *Proteomics* 2006, **6**(9):2759-2771.

7. Lu P, Vogel C, Wang R, Yao X, Marcotte EM: **Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.** *Nat Biotechnol* 2007, **25**(1):117-124.

8. Kuster B, Schirle M, Mallick P, Aebersold R: **Scoring proteomes with proteotypic peptide probes**. *Nat Rev Mol Cell Biol* 2005, **6**(7):577-583.

9. Richard E. Higgs MDK, Valentina Gelfanova, Jon P. Butler, and John E. Hale: **Comprehensive Label-Free Method for the Relative Quantification of Proteins from Biological Samples**. **Journal of Proteome Research** 2005, **4**:1442-1450.

10. Washburn MP, Ulaszek RR, Yates JR, 3rd: **Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology**. *Anal Chem* 2003, **75**(19):5054-5061.

11. Nanduri B, Lawrence ML, Boyle CR, Ramkumar M, Burgess SC: **Effects of subminimum inhibitory concentrations of antibiotics on the Pasteurella multocida proteome**. *J Proteome Res* 2006, **5**(3):572-580.

12. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF: **Detecting differential and correlated protein expression in label-free shotgun proteomics**. *J Proteome Res* 2006, **5**(11):2909-2918.

13. Kawashima S, Kanehisa M: **AAindex: amino acid index database**. *Nucleic Acids Res* 2000, **28**(1):374.

14. Witten IH, Frank E: Data Mining: **Practical Machine Learning Tools and Techniques**, 2 edn. San Francisco: Morgan Kaufmann; 2005.

15. Burgess SC, Young JR, Baaten BJ, Hunt L, Ross LN, Parcells MS, Kumar PM, Tregaskes CA, Lee LF, Davison TF: **Marek's disease is a natural model for lymphomas overexpressing Hodgkin's disease antigen (CD30)**. *Proc Natl Acad Sci* U S A 2004, **101**(38):13879-13884.

16. Burgess SC, Davison TF: **Identification of the neoplastically transformed cells in Marek's disease herpesvirus-induced lymphomas: recognition by the monoclonal antibody AV37**. *J Virol* 2002, **76**(14):7276-7292.

17. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: **Protein Identification and Analysis Tools on the ExPASy Server**. In: *The Proteomics Protocols Handbook*. Edited by Walker JM: Humana Press; 2005.

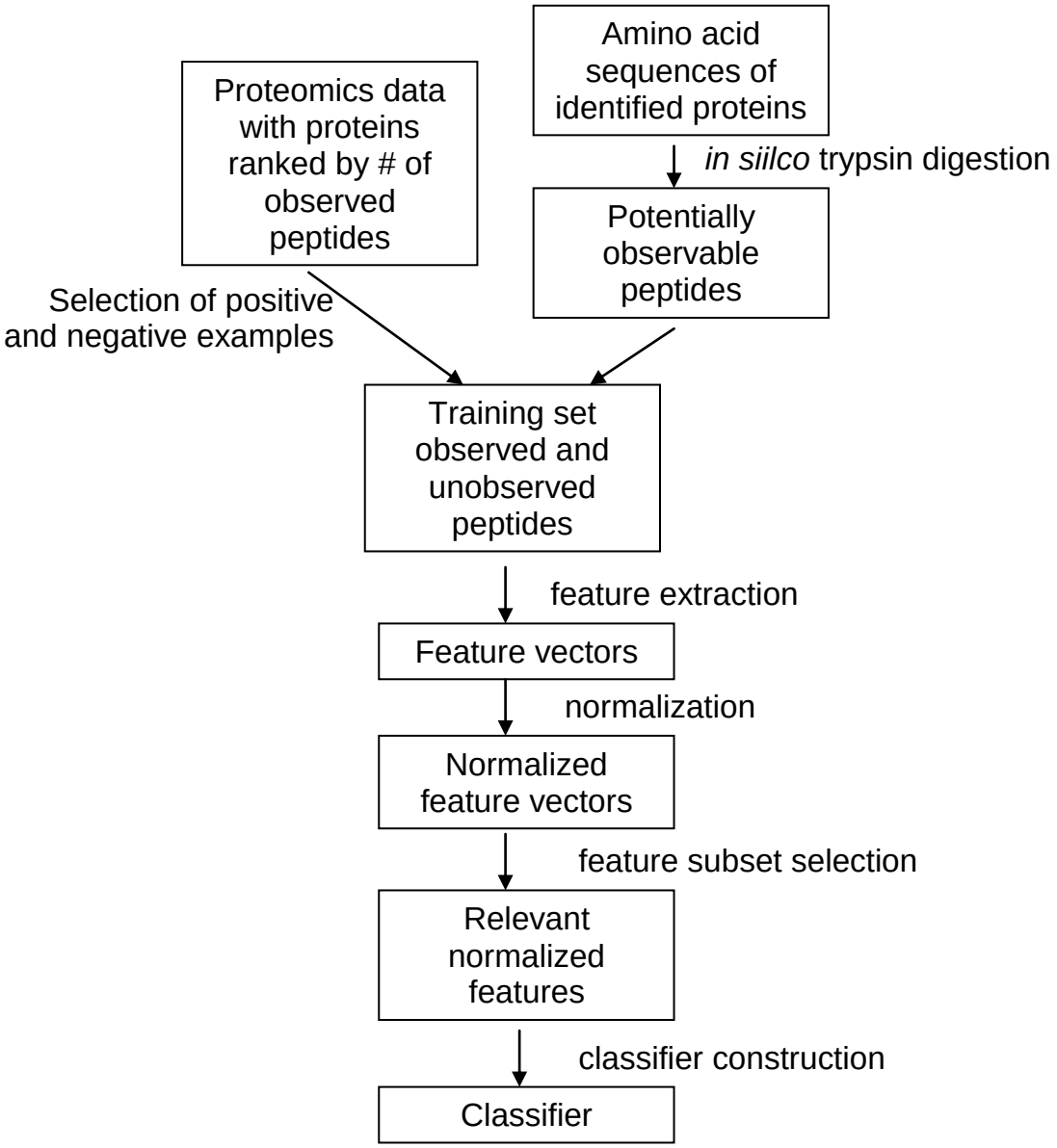


Figure 2.1

Classifier construction process

TABLE 2.1

A LIST OF INITIAL FEATURES USED FOR CLASSIFIER CONSTRUCTION IN ADDITION TO AAINDEX FEATURES.

Feature Subset 1
Length of peptide
Net charge of peptide
Positive charge
Negative charge
Isoelectric point
Molecular weight
Hydropathicity
Count of each amino acid (20 features)
Percent composition of each amino acid (20 features)
Percent polar amino acids
Percent positive amino acids
Percent negative amino acids
Percent hydrophobic amino acids

NOTE: A feature selection procedure is used to reduce dimensionality prior to classifier construction.

TABLE 2.2

DESCRIPTION OF FEATURES SELECTED FOR THE CLASSIFIERS BUILT FOR THE TWO DATASETS.

Avian Bursa Dataset
Number of prolines
Percent glycine
Percent alanine
Percent leucine
Percent polar amino acids
Percent hydrophobic amino acids
Percent positive amino acids
Percent negative
Size (Dawson, 1972)
Optimized transfer energy parameter (Oobatake et al., 1985)
Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)
Transfer free energy from oct to wat (Radzicka-Wolfenden, 1988)
Information measure for C-terminal turn (Robson-Suzuki, 1976)
Amphiphilicity index (Mitaku et al., 2002)
Hodgkin's Lymphoma Model Dataset
Number of cysteine
Signal sequence helical potential (Argos et al., 1982)
Transfer free energy to surface (Bull-Breese, 1974)
Normalized relative frequency of alpha-helix (Isogai et al., 1980)
Normalized relative frequence of double bend (Isogai et al., 1980)
Distance between C-alpha and centroid of side chain (Levitt, 1976)
Retention coefficient in NAH2PO4 (Meek-Rossetti, 1981)
Interior composition of amino acids intracellular proteins (Fukuchi-Nishikawa, 2001)
Linker propensity from 1-linker dataset (George-Heringa, 2003)

TABLE 2.3

10-FOLD CROSS-VALIDATION ACCURACY BY CLASS FOR NEURAL NETWORKS GENERATED FOR TWO DATASETS.

Class	True positive rate	False positive rate	Precision	Recall	ROC Area
Avian Bursal Dataset					
Not observed	0.80	0.19	0.81	0.80	0.87
Observed	0.82	0.20	0.80	0.82	0.87
Hodgkin's Lymphoma Model Dataset					
Not observed	0.66	0.22	0.75	0.66	0.80
Observed	0.78	0.34	0.70	0.78	0.80

TABLE 2.4

ACCURACY BY CLASS FOR NEURAL NETWORKS GENERATED USING ONE DATASET AS THE TRAINING SET AND THE OTHER DATASET FOR TEST DATA.

Class	True positive rate	False positive rate	Precision	Recall	ROC Area
Avian Bursal Dataset training set, Hodgkins Lymphoma test set					
Not observed	0.71	0.46	0.61	0.71	0.70
Observed	0.54	0.29	0.66	0.54	0.70
Hodgkin's Lymphoma Model Dataset training set, Avian Bursa test set					
Not observed	0.81	0.41	0.81	0.73	0.73
Observed	0.59	0.19	0.59	0.66	0.73

TABLE 2.5

NUMBER OF TRYPTIC PEPTIDES PREDICTED TO BE OBSERVABLE FOR SELECTED PROTEINS FROM THE TWO DATA SETS.

Protein GI Number	Num tryptic peptides (>= 6 aa)	Num tryptic peptides observed	Percent amino acid coverage	Number predicted detectable	Percent predicted detectable	Percent amino acid coverage of detectable
Avian bursa data set						
5902793	20	2	10	9	45	33
119359	50	5	9	15	30	21
128413	16	2	11	3	18	14
2119012	7	2	28	3	43	17
17025728	16	2	6	7	44	20
122000	6	4	33	0	0	0
1762374	7	1	23	2	29	21
1172808	13	1	6	4	30	19
7512219	44	1	2	11	25	34
104697	9	2	22	4	44	30
118106991	12	0	0	0	0	0
Hodgkin's lymphoma model data set						
479367	34	1	3	5	15	11
729629	18	2	14	11	61	43
899264	13	1	10	4	31	21
63544	48	2	2	6	13	15
50750413	38	3	11	8	21	25
45433516	26	0	0	0	0	0
46048702	14	0	0	0	0	0
125745137	9	0	0	0	0	0
125745114	9	0	0	0	0	0
45433516	26	0	0	0	0	0

NOTE:1) For the avian bursa dataset, 10 randomly selected observed proteins (in white) and the DR3 protein that was expected but not observed (in yellow). 2) For the Hodgkin's lymphoma model dataset, 5 proteins that were observed in the pathway under consideration and 5 (in yellow) that had been observed using other methods in previous experiments but not observed in this dataset.

CHAPTER III

PREDICTION OF CELL PENETRATING PEPTIDES BY SUPPORT VECTOR MACHINES

Abstract

Cell penetrating peptides (CPPs) are those peptides that can transverse cell membranes to enter cells. Once inside the cell, different CPPs can localize to different cellular components and perform different roles. Some generate pore-forming complexes resulting in the destruction of cells while others localize to various organelles. Use of machine learning methods to predict potential new CPPs will enable more rapid screening for applications such as drug delivery. We have investigated the influence of the composition of training datasets on the ability to classify peptides as cell penetrating using support vector machines (SVMs). We identified 111 known CPPs and 34 known non-penetrating peptides from the literature and commercial vendors and used several approaches to build training data sets for the classifiers. Features were calculated from the datasets using a set of basic biochemical properties combined with features from the literature determined to be relevant in the prediction of CPPs. Our results using different training datasets confirm the importance of a balanced training set with approximately equal number of positive and negative examples. The SVM based classifiers have greater classification accuracy than previously reported methods for the prediction of CPPs, and

because they use primary biochemical properties of the peptides as features, these classifiers provide insight into the properties needed for cell-penetration. To confirm our SVM classifications, a subset of peptides classified as either penetrating or non-penetrating was selected for synthesis and experimental validation. Of the synthesized peptides predicted to be CPPs, 100% of these peptides were shown to be penetrating.

Introduction

Cell penetrating peptides (CPPs), also referred to as "Trojan" peptides, protein transduction domains, or membrane translocation sequences, are typically hydrophobic linear arrangements of 8-24 amino acids able to cross the lipid bi-layer membrane that serves as the cell's outer barrier and gain access to the interior of the cell and its components [1]. Penetratin, an Antennapedia derived peptide, and the HIV derived Tat peptide were some of the first commonly studied CPPs, and along with transportan peptides (derived from galanin receptor ligand proteins), make up three major families of CPPs. The remainder of CPPs are classified in a fourth, miscellaneous family [1].

Initially, cellular uptake of CPPs was believed to be through endocytosis or protein transporters, but some evidence suggested the mechanism may involve direct transport through the lipid bi-layer of the cell, which takes into account the hydrophobic properties of most of these peptides [2]. The current view is that CPP internalization is accomplished predominantly by endocytosis [3]. Historically, both flow cytometry and fluorescence microscopy have been used to study the uptake of CPPs into cells. Care must be used with these methods to avoid artifacts because traditional methodologies for

these techniques can incorrectly show a high concentration of CPPs localizing to the cell nucleus or a higher than actual concentration of CPPs being taken into the cell [2].

Cell penetrating peptides capable of transporting other active molecules inside the cell have the potential to serve as drug delivery peptides. Drug delivery peptides and CPPs allow researchers to probe the mechanisms of peptide transport across a lipid bi-layer membrane and may allow customizable drug therapies for differing types of cells. Although there is some controversy regarding CPPs as drug delivery systems because of their lack of specificity for cell type, the general consensus among researchers is that both general CPPs and cell-specific CPPs will be developed into effective drug delivery systems in the future [4, 5].

A classification system that can determine whether or not a unique peptide sequence can serve as a CPP, and thus possibly be a potential drug delivery peptide, can enable researchers to quickly screen candidate molecules for their potential viability for use in a customizable drug delivery regime.

Much of the previous work in the prediction of CPPs has involved the use of a set of composite features assembled from primary biochemical properties through the use of principal component analysis [6-8]. These composite features, or *z*-scores, consist of a numerical value and an associated range. To predict cell-penetrating capability of a candidate peptide, the *z*-scores are computed for the peptide, and, if the *z*-scores fall within the range of known CPP *z*-scores, the peptide is classified as cell-penetrating [7, 8]. While this method has a high accuracy (>95% correct prediction of novel CPPs) for generating novel CPPs [7], it performs rather poorly (68% correct prediction) when

trying to distinguish known non-penetrating peptides that are closely related to known CPPs [8] and yields little information about exactly which biochemical properties contribute to the difference between these two classes. More recent work examines the use of quantitative structure-activity relationship (QSAR) derived features to predict penetration potential. The training process iteratively removes sequences that are difficult to classify and thus the classification accuracies reported are biased [9]. Further research into this topic is necessary to allow potential drug delivery peptides to be rapidly screened for usefulness.

Using the basic biochemical properties of peptides as features instead of the widely used composite *z*-scores can potentially provide more insight into the differences between the class of CPPs and non-penetrating peptides when coupled with the use of a machine learning classifier such as a support vector machine. Additionally, once trained, these machine learning classifiers can then be used for rapid screening of candidate CPPs prior to their synthesis. This study examines the available information on known CPPs and their non-penetrating analogs in order to compile datasets used for training and testing of support vector machine classifiers using primary features derived from biochemical properties of each peptide and evaluates the accuracy of these classifiers. An experimental validation study was performed to determine the effectiveness of these classifiers using an avian tissue culture system.

Results and Discussion

The goal of this study was to develop a machine learning approach for rapid screening of potential CPPs. We use features representing primary biochemical properties directly rather than using a transformation such as PCA that combines multiple features into a single composite feature as reported by others [6-8]. In addition, we have investigated the best approach for constructing training datasets when there is a large disparity in the number of positive and negative examples. Previous research has shown that unbalanced datasets are problematic when constructing classifiers [10]. We first identified known CPPs and known non-penetrating peptides from the literature to serve as positive and negative examples and calculated a number of primary biochemical properties for each of these peptides. We then explored a number of different approaches for addressing the problem of unbalanced datasets and evaluated classification accuracy with the different approaches. A wrapper based feature selection method was utilized to reduce the number of features needed for classification while providing insight into the biochemical properties necessary to distinguish CPPs from non-CPPs. We have used support vector machine classifiers because of their ability to linearly separate classes in a high dimensional feature space. Classifier accuracy on our training sets was assessed using 10-fold cross validation and then each classifier was tested again using the unbalanced test set assembled from the literature. In order to experimentally validate these results, a dataset of 250 peptides was created using a 0^{th} order Markov model based on the predicted chicken proteome [11], and these peptides were classified as either penetrating or non-penetrating by our classifier. Subsets of both predicted penetrating

and predicted non-penetrating peptides were selected from these classification results and were synthesized. Experimental validation of cell penetration capability was then determined using fluorescence microscopy and the quantitative uptake of peptides shown to be penetrating was performed.

Dataset Construction Approaches

Because of the sensitivity of classifiers to unbalanced classes [10], our first challenge was to generate datasets for training and testing. A set of 111 known CPPs were identified from the literature [7, 8, 12]. However, only 34 negative examples could be found and many of these are analogs of known CPPs [7, 8]. Unbalanced datasets present a number of different problems for machine learning methods [10]. When only a comparatively small number of examples are available for one class, the machine learning algorithm will not have sufficient information to learn a function to distinguish the classes. Reporting of classification accuracy is also impacted by unbalanced datasets. For example, if a dataset of 100 peptides contains 80 CPPs and 20 non-CPPs, a classification accuracy of 80% can be obtained by classifying all peptides as positive. Most previous work in CPP prediction has ignored this problem [8, 9].

We designed an experiment to investigate the effect of unbalanced datasets on CPP prediction and to find methods to address the problem to evaluate classifier accuracy with precision. For the CPP prediction problem, there are many more positive examples than negative examples available. Five different approaches were used to generate training datasets for investigating this issue:

1. *Unbalanced:* Composed of 34 known negative examples and 111 known positive examples.

2. *Balanced with random peptides as negative examples*. 111 random peptides were generated using a 0^{th} order Markov chain based on the chicken proteome and combined with 111 known positive examples. All random peptides were assumed to be non-penetrating. This approach is based on the assumption that the probability of randomly generating a CPP sequence is very small.

3. *Balanced with biological peptides as negative examples*. All chicken peptides of length 12-26 AA were downloaded from NCBI and a sample of 111 was drawn without replacement. All were assumed to be non-penetrating. This approach assumes that most biological peptides are non-CPP and the probability of drawing a CPP from this set is extremely low.

4. *Balanced by sampling known negatives*. Random sampling with replacement from the 34 known negatives was used to yield a set of 111 negative examples that was combined with the 111 positive examples.

5. *Balanced by sampling known positives*. Random sampling with replacement from the 111 known positive examples to yield a set of 34 positive examples that was combined with the 34 known negative examples.

Classifier Performance

The performance of all classifiers on the training data sets is based on 10-fold cross validation. The confusion matrices for classifiers trained using datasets based on approaches 1-4 are shown in Table 3.1 and the classifier statistics are shown in Table 3.2. The classifier trained on the unbalanced dataset (111 positive examples and 34 negative

examples) has a classification accuracy of only 75.86% compared to the naïve approach of classifying all examples as positive which would result in a classification accuracy of 76.55%. The results for this dataset in Table 3.1 show that the resulting classifier predicts almost all examples to be positive. This highlights the problems encountered when using an unbalanced dataset. The classifier cannot distinguish positive and negative examples because the dataset contains so many more positive examples than negative examples and because many of the negative examples are analogs of the positives.

The classifiers trained using both the dataset balanced with random peptides for negatives (approach 2) and with biological peptides for negatives (approach 3) had classification accuracies of 95.95% and 94.14% respectively, indicating that both classifiers exhibit a high degree of accuracy in discriminating between known cell-penetrating peptides and randomly generated or biological peptides assumed to be negative. The confusion tables for these classifiers on the training data sets (Table 3.1) show that most of the mistakes are false negatives (CPPs incorrectly classified as non-CPPs). The weakness of these training approaches is that some of the assumed negative examples may in fact be cell penetrating and known non-cell penetrating analogs of CPPs were not used as negative examples. When we used these trained classifiers to evaluate the known non-penetrating cell penetrating analog peptides (our unbalanced test data set) these classifiers obtained accuracies of 80.69% and 79.31% respectively. For both classifiers, approximately one third of the known non-penetrating peptides are classified as cell-penetrating. Most of the mistakes made by these two classifiers on the test data

seem to be false positives, that is classifying a peptide with no cell penetrating potential as a CPP, and this classification of known non-penetrating cell penetrating analogs demonstrates that while these classifiers are very accurate distinguishing the features strongly predictive of cell penetrating potential from the vast majority of non-penetrating peptides, the features used for classification do not serve to distinguish between peptides more similar to CPPs that do not penetrate and those peptides that can act as CPPs.

The classifier trained on the data set constructed using approach 4 (random sampling with replacement from the known negative examples) has a classification accuracy of 88.74% on the training data set when evaluated with 10-fold cross validation. When compared to the classification accuracy of the dataset generated using the unbalanced dataset, these results show that it is possible to classify a set of CPPs and a set of known non-penetrating peptides using our SVM based method when care is used to construct balanced datasets. Table 3.2 shows that 60% of the errors are false positives (non-CPPs incorrectly classified as CPPs). When we evaluated the unbalanced test set on this classifier, an accuracy of 91.72% was obtained. The classifiers trained on the smaller datasets using approach 5 have an average classification accuracy of 78.82% using 10-fold cross validation.

Approach 2 using randomly selected biological peptides as the negative examples gives the best 10-fold cross validation accuracy while approach 4 with random selection from the negative examples gives the best accuracy for the unbalanced training set. This suggests use of a two step process for screening. In the first step, a classifier trained with random biological peptides as the negative examples would be used for preliminary bulk

screening. As a second step, peptides predicted to be CPP in step 1 would be screened by a classifier trained using approach 4 that is more accurate in distinguishing non-penetrating analogs from CPPs. Approach 4 also provides more insight into the rational design of novel CPP analogs as the negative examples used in this approach are generally constructed by the modification of a known CPP sequence.

In Hällbrink et al. (2005), the authors describe a method of CPP prediction based on scoring a candidate peptide according to *z*-score descriptors, features compiled through PCA, and report an 84.05% accuracy in the prediction of 53 CPPs and 16 non-functional CPP analogs [7]. A follow-up to this study, utilizing both more known CPPs (65) and more non-functional CPP analogs (20), reports a 68% prediction efficiency using the same *z*-score descriptor based prediction method [8]. More recently, these *z*-score descriptors were utilized alongside quantitative structure-activity relationship features in an artificial neural network (ANN) to predict cell penetrating potential for a set of 101 peptides (77 CPPs, 24 non-penetrating CPP analogs) and report a classification accuracy of 83% for the general ANN model constructed [9]. However, it should be noted that the data set utilized is composed of unbalanced classes, and an accuracy of 76.24% can be achieved by classifying every peptide encountered as a CPP. A comparison of these previously published prediction methods and our approach is presented in Table 3.3. The models constructed using our approaches and their high classification accuracies indicate that using the primary biochemical properties of a peptide as features instead of synthesized feature values compiled using PCA allows for a more informative analysis of which properties determine whether a given peptide is cell-

penetrating. Our approach also allows predictive models constructed on training sets to be used for more rapid and elucidative screening of cell-penetrating potential than previous predictive methods based on verifying whether a given peptide falls within some average range of composite features.

For each classifier constructed, feature selection was conducted using a scatter search approach through feature space [13] where the "wrapped" classifier was the same type of SVM used for classifier construction. The classifier is a sequential minimal optimization SVM [14] using the Pearson Universal Kernel [15]. Table 3.4 lists the features selected for datasets 1-4 above. Because the number of training/testing samples for dataset 5 was so small, we generated ten different datasets using this approach. The features selected from these ten datasets are listed in Table 3.5. The features selected for the datasets constructed using approaches 1-5 contain a number of properties previously shown to aid in the prediction of CPPs. These include net charge, positive charge, negative charge, the net donated hydrogen bonds, and the water-octanol partition coefficient. The low number of features selected for the datasets constructed using approach 5 indicates over-fitting of these small datasets by the classification algorithm. Therefore our detailed examination of features selected focused on datasets generated using approaches 1-4. The primary amino acid composition features, the number of a given amino acid and the percent a given amino acid contributes to the whole peptide sequence, indicates no predictive function arising from the non-polar amino acids leucine and isoleucine, the polar amino acid glutamine, and the negatively charged amino acid glutamate. At least one of the amino acid composition features was selected for the

remaining amino acids, with the most notable of these being the positively charged amino acids lysine, arginine, and histidine, and the negatively charged amino acid aspartate. In addition, the group of aromatic amino acids were selected to a notable degree, and the presence of some aromatic amino acids in the peptide sequence has been previously reported to be required for cell-penetrating potential [16].

Validation study

To experimentally validate our feature selection methodology and classifiers, 250 random peptides were generated using a 0^{th} order Markov model based on the chicken predicted proteome and were classified as penetrating or non-penetrating using the classifier trained on the dataset constructed using random peptides as negative examples. From these classifications, four peptides predicted to be cell-penetrating and two peptides predicted to be non-penetrating were selected for synthesis and FITC-labeling along with three known cell penetrating peptides used for positive controls, three peptides consisting respectively of only polar amino acids, only non-polar amino acids, and only of mixed polar and non-polar amino acids to serve as negative controls. In addition, a known non-penetrating peptide (TP13, a transportan analog [16]) was selected for synthesis to serve as a minor validation for our set of known non-penetrating peptides.

Cellular Internalization Microscopy Array of FITC-Labeled Peptides

The uptake of synthesized FITC-labeled peptides was examined using an avian system to validate both our wrapper based feature selection methodology and SVM-based approach to predicting CPPs. The results of our fluorescence microscopy analysis are

shown in Figure 3.1. All peptides predicted to be cell-penetrating (Peptide-1 through Peptide-4) by our classifier were confirmed to be cell-penetrating. Of our two negative predictions, Peptide-5 was confirmed to be a non-penetrating peptide while Peptide-6 was shown to traverse cellular membranes. TP13, a CPP analog previously shown to be non-penetrating in Bowes' melanoma cells is clearly cell-penetrating peptide in our avian model.

Uptake Quantification of FITC-Labeled Peptides

To evaluate the relative uptake of our synthesized peptides and to provide a secondary confirmation of the fluorescence microscopy results, a quantitative uptake study was conducted using both quail SOgE cells and chicken embryonic fibroblasts. The results of the quantitative uptake study are shown in Figure 3.2. Peptides 1-4 were shown to be CPPs, while Peptide-5 was correctly predicted to be non-penetrating. Peptide-6, which was predicted to be non-penetrating, was shown to traverse the membranes of both CEF and SOgE cells. TP13, previously shown to be non-penetrating in melanoma cells, is again shown to have penetrated both CEF and SOgE cells to a high degree relative to both our positive controls and our predicted cell-penetrating peptides. TP13 was chosen as a non-penetrating CPP analog based on its non-CPP classification in a study examining the effects of deletion on a known CPP, transportan (TP) [16]. TP13 was created by a deletion from the N-terminus and middle of the TP molecule and these deletions abolished the internalization of TP13 into Bowes' melanoma cells. All transportan-derived peptides that internalized during the original TP analog study contained tyrosine and 3 positive charges in their sequences, while those peptides without

tyrosine or one positive charge in the C-terminal portion of the peptide did not internalize [16]. TP13 contains tyrosine and 3 positive charges, meeting the criteria outlined by the original study for penetration and both our fluorescent microscopy data and quantitative fluorescent uptake data indicates that it does penetrate both SOgE cells and CEF cells.

Peptide-6 (HSPIIPLGTRFVCHGVT) was predicted to be a non-CPP by our classifier, but was shown to internalize into both SOgE and CEF cells experimentally both by fluorescence microscopy and the quantitative uptake studies. This peptide contains 3 positively charged amino acids along with phenylalanine. The Sommets, *et al*. study examining TP and its derivatives states that all their peptides with 3 positive charges and tyrosine internalized, and as phenylalanine only lacks the hydroxyl group of the tyrosine molecule, this could contribute to the internalization of Peptide-6. The positive examples in our training data contain predominantly arginine and lysine as positive residues, while this peptide contains two histidine residues.

Our research shows that using the primary biochemical properties of peptides as features instead of composite features determined through the use of PCA can provide both more informative features and higher classification accuracies when using support vector machines for the classification of a given peptide as cell-penetrating. The lack of a comprehensive and coherent database of cell-penetrating peptide data for bioinformatics analysis has been noted previously [8], and the majority of CPP studies have been conducted using a variety of different cell lines and detection techniques, making it difficult to unify these results. Our results showing that a previously reported non-penetrating analog of transportan is a CPP in our avian system confirms the need for a

large dataset of biologically confirmed positive and negative examples from a single biological system using a single detection methodology. Until such a resource is available, the predictive capability of classifiers is difficult to assess. Our results also show that there may be classes of peptides that act as CPPs in a variety of cells and others that are more specialized. Therefore, peptides designed to target delivery to specific cells and tissues of interest should be screened using a variety of cell lines. Additionally, our results indicate there may be positional preference for certain types of amino acids such as positive charges and aromatic. Further research should examine the effects of these positional effects.

Materials and Methods

Data Set Compilation Strategy

A database of cell-penetrating peptides was constructed from the literature and from commercial vendor product lines [7, 8, 12]. A total of 111 cell-penetrating peptide (CPP) sequences were identified and used to create a database of positive examples (Table 6) [7, 8, 12]. The average amino acid lengths of these CPPs ranged from 12 to 26. Because very few peptides have been experimentally validated to be non-penetrating, it was more challenging to construct a database of negative examples. Five different strategies were used. Because our experimental system is avian, we have used the composition of the chicken proteome as the basis for two of our datasets. Previous research has demonstrated the importance of using a balanced training sets where there

are approximately equal numbers of positive and negative examples [10]. Our strategies are listed below:

1. *BALANCED WITH RANDOM PEPTIDES*: The set of 111 know CPPs was balanced with a set of 111 peptides constructed using a 0^{th} order Markov chain derived from the IPI chicken proteome (ipi.CHICK.v3.56 [11]) . The peptide lengths were uniformly distributed in the range 12-26. We assume that there is a very low probability that randomly generated peptides would be cell penetrating.

2. *BALANCED WITH BIOLOGICAL PEPTIDES*: The set of 111 know CPPs was balanced with randomly selected biological peptides. A set of 411 chicken peptides from NCBI with lengths in the range 12-26 was downloaded. Subsets of 111 peptides were selected randomly without replacement to provide multiple balanced datasets. This dataset provides a set of positive examples of known CPPs and assumed negative examples of biological peptides of the same relative molecular size. We assume that most naturally peptides are not cell penetrating.

3. *UNBALANCED USING ONLY KNOWN POSITIVES:* A set of 34 known non-penetrating cell penetrating peptide analogs and peptide hormones previously used as negative examples was constructed from a search of the literature and are listed in Table 3.7 [7, 8]. This dataset provides a set of known cell-penetrating positive examples and a set of non-penetrating peptides that have been experimentally shown not to traverse cellular membranes.

4. *BALANCED BY SAMPLING KNOWN NEGATIVES*: In order to produce a balanced dataset of both known non-penetrating peptides and known CPPs a set consisting of all 111 known cell penetrating peptides and 111 known non-penetrating cell penetrating analogs was constructed by selecting with replacement from the set of 34 known non-penetrating analogs .

5. *BALANCED BY SAMPLING KNOWN POSITIVES*: Subsets of the known CPPs of size 34 were selected with replacement and combined with the 34 known non-penetrating cell penetrating analogs to create ten balanced subsets.

Feature Construction and Normalization

For each dataset, we generate a set of basic biochemical properties of each peptide (e.g. mass, size, charge, secondary structure, etc) and other features previously shown to be useful in the prediction of CPPs (e.g. steric bulk and net donated hydrogen bonds) [8]. The full list of the initial 61 features is shown in Table 3.8. We use these features directly in our machine learning algorithm rather than using composite features such as features derived by principle component analysis [8, 17]. We feel this approach will be more informative in the rationale design of CPPs cell penetrating peptides. Because the data values for each feature within a dataset vary greatly, NV normalization was used to scale the numeric range of all features in the range [0, 1] [18].

Machine Learning Software

The WEKA Machine Learning Toolkit Version 3.6.1, a freely available software package containing a number of machine learning algorithms for data mining, was used for feature selection, classifier construction, and classifier evaluation [19].

Feature Selection

We conducted feature selection to reduce the dimensionality of the feature vectors. Empirical evaluation of a number of different feature selection methods was conducted and the best performance was obtained using a wrapper-based method. The wrapper-based method uses a parallel scatter search algorithm [13] to evaluate feature subsets based on classifier performance. Scatter search is an evolutionary algorithm, but unlike other evolutionary algorithms (e.g. genetic algorithms), the search for a local

optimum is guided through the use of a reference set that acts to intensify and diversify the resulting features [13]. Local searches of features generated from the reference set are conducted, and informative and diverse features from these local searches are used to update the reference set until a terminating condition is met [13].

Classifier Construction

Our classifier is a support vector machine (SVM) trained via a sequential minimal optimization (SMO) algorithm used in conjunction with the Pearson VII universal kernel [14, 15]. SVMs are supervised learning classifiers generally used for solving two class problems, and in their simplest form can be thought of as a classifier separating two classes mapped onto a 2-dimensional plane by generating a line through the plane that optimizes the distribution of each class on either side of the line [14]. The SMO algorithm is a modification to the original SVM learning algorithms that replaces a numerical quadratic programming step with an analytical quadratic programming step, allowing the algorithm to spend a greater portion of time on the decision function instead of the quadratic programming step. This greatly increases the speed of the SVM for classification and allows scaling for large datasets [14]. We chose to utilize SMO-based SVM classifiers because of their speed and performance for our two class problem of determining if given peptide is cell-penetrating or non-penetrating. A kernel function used in conjunction with an SVM allows the classifier to examine non-linear relationships between features by mapping the initial non-linear features into a highly dimensional space where the solution can be represented by a linear classification [15]. We chose the Pearson VII universal kernel (PUK) for our SMO-based SVM because

PUK has been shown to provide either equal or better mapping than traditional SVM kernels, while serving as a robust and generic alternative to other kernel functions [15]. Accuracy for all classifiers was evaluated using 10-fold cross-validation.

Peptide Synthesis

A 0^{th} order Markov chain based on the amino acid frequency of the IPI Chicken Proteome (ipi.CHICK.v3.56) [11] was used to generate 250 peptides. The classifier trained on our biologically based random peptide dataset was then used to classify each of these peptides. From these classification results, four peptides predicted to be cell penetrating and two peptides predicted to be non-cell penetrating were selected for synthesis and experimental validation. In addition, three peptides known to be cell-penetrating (HIV-Tat [20], Antennapedia [21], and Pep-1 [22]) were chosen to be positive experimental controls. Three other peptides, one of all polar amino acids, one of all non-polar amino acids, and one of a mix of polar and non-polar amino acids, were chosen as negative experimental controls because their lack of charged and aromatic R-groups make it unlikely they would cross a cellular membrane. One peptide (TP13 [8, 16]) was randomly selected for synthesis from the list of known non-penetrating cell penetrating peptide analogs. All peptides selected for synthesis are shown in Table 3.9. Peptides were synthesized (>95% purity) and N-terminally labeled with FITC, a fluorescent tag, by Biomatik. During the peptide synthesis, one of our chosen negative controls, negative-2 (GLALLGIAVAILVVL-NH_2) was unable to be synthesized to our desired purity levels due to insolubility issues and is not considered further. The

lyophilized peptides were reconstituted using 1 mL of 4:1 dd H_2O sterile filtered 0.45 µm and acetonitrile (EMD OmniSolv).

Tissue Culture

Two avian cell lines, Quail SOgE muscle cells [23] and a primary culture of Chicken embryonic fibroblasts (CEF), were grown in tissue culture flasks in Dulbecco's minimal essential medium containing 10% fetal bovine serum with penicillin (200 IU/mL), streptomycin (200 µg/mL), amphotericin B (0.5 µg/mL) (MP Biomedicals), and non-essential amino acids at 37°C in a 5% CO_2 atmosphere.

Quantitative Uptake Analysis

Approximately 100,000 cells per well (both CEFs and SOgEs) were plated onto 12-well tissue culture plates approximately 2 days prior to the experiment and allowed to reach confluency. The cells were changed to serum free media and incubated for 60 minutes prior to experimentation. The cells were then washed with two 1 mL washes of PBS, after which they were exposed to 300 µL of 10 µM peptide in serum free media for 30 minutes, with three replicates per peptide per cell line. The cells were then washed with two 1 mL washes of PBS, and lightly trypsinated to facilitate their detachment from the plate. Cells were then lysed with 250 µL of 0.1% Triton-X in PBS at 4° C for 10 minutes. A 100 µL aliquot of the cell lysate and a 100 µL aliquot of the 10 µM peptide in serum free media were pipetted onto a 96-well plate. Fluorescence was measured on a Dynex Fluorolite 1000 plate reader at 485/530nm. The samples were compared to the

fluorescence of the added amount of peptide and *t*-tests ($p > 0.05$) were performed for each experimental sample against an untreated control.

Cellular Internalization Microscopy Array of FITC-Labeled Peptides

The SOgE cells were seeded onto glass tissue microscopy slides (approximately 50,000 cells/well), and allowed two days to reach confluency. The cells were changed to serum free media and incubated for 60 minutes prior to experimentation. The cells were then washed with two 1 mL washes of PBS, after which they were exposed to 300 µL of 10 µM peptide in serum free media for 30 minutes. The cells were then washed with two 1 mL washes of PBS, and then fixed using UltraCruz™ Mounting Medium (Santa Cruz Biotechnology) containing a DAPI nuclear stain. The fluorescence was examined using a Nikon Eclipse TE2000-U Inverted Research Microscope with the MetaMorph microscopy imaging software.

LITERATURE CITED

1. Kilk K: **Cell-penetrating peptides and bioactive cargoes. Strategies and mechanisms.** Stockholm: Stockholm University; 2004.

2. Richard JP, Melikov K, Vives E, Ramos C, Verbeure B, Gait MJ, Chernomordik LV, Lebleu B: **Cell-penetrating peptides. A reevaluation of the mechanism of cellular uptake**. *J Biol Chem* 2003, **278**(1):585-590.

3. Yanagisawa K, Shyr Y, Xu B, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S *et al*: **Proteomic patterns of tumor subsets in non-small-cell lung cancer**. *The Lancet* 2003, **362**(9382):433-439.

4. Schwartz JJ, Zhang S: **Peptide-mediated cellular delivery**. *Curr Opin Mol Ther* 2000, **2**(2):162-167.

5. Vives E: **Present and future of cell-penetrating peptide mediated delivery systems: "is the Trojan horse too wild to go only to Troy?"**. *J Control Release* 2005, **109**(1-3):77-85.

6. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids**. *J Med Chem* 1998, **41**(14):2481-2491.

7. Hallbrink M, Kilk K, Elmquist A, Lundberg P, Lindgren M, Jiang Y, Pooga M, Soomets U, Langel U: **Prediction of Cell-Penetrating Peptides**. *International Journal of Peptide Research and Therapeutics* 2005, **11**(4):249-259.

8. Hansen M, Kilk K, Langel U: **Predicting cell-penetrating peptides**. *Adv Drug Deliv Rev* 2008, **60**(4-5):572-579.

9. Dobchev DA, Mager I, Tulp I, Karelson G, Tamm T, Tamm K, Janes J, Langel U, Karelson M: **Prediction of Cell-Penetrating Peptides Using Artificial Neural Networks**. *Curr Comput Aided Drug Des*, **2010**:6.

10. Provost F: **Machine Learning from Imbalanced Data Sets 101**. In: *AAAI Workshop on Learning from Imbalaced Data Sets: 2000*: The AAAI Press; 2000: 1-3.

11. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments**. *Proteomics* 2004, **4**(7):1985-1988.

12. **Cell Permeable Peptides (CPP) / Drug Delivery Peptides** [http://www.anaspec.com/products/productcategory.asp?id=158]

13. Lopez F, Torres M, Batista B, Perez J, Moreno-Vega J: **Solving feature subset selection problem by a Parallel Scatter Search**. *European Journal of Operational Research* 2006(169):477-489.

14. Platt J: **Fast Training of Support Vector Machines using Sequential Minimal Optimization**. In: *Advances in Kernel Methods: Support Vector Learning.* Edited by Scholkopf B, Burges C, Smola A. Cambridge, MA: MIT Press; 1999: 185-208.

15. Ustun B, Melssen W, Buydens L: **Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel**. *Chemometrics and Intelligent Laboratory Systems* 2005(81):29-40.

16. Soomets U, Lindgren M, Gallet X, Hallbrink M, Elmquist A, Balaspiri L, Zorko M, Pooga M, Brasseur R, Langel U: **Deletion analogues of transportan**. *Biochim Biophys Acta* 2000, **1467**(1):165-176.

17. Hallbrink M, Kilk K, Elmquist A, Lundberg P, Lindgren M, Jiang Y, Pooga M, Soomets U, Langel U: **Prediction of Cell-Pentrating Peptides**. *International Journal of Peptide Research and Therapeutics* 2005, **11**(4):10.

18. Liu H, Li J, Wong L: **A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns**. *Genome Informatics* 2002, **13**:51-60.

19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I: **The WEKA Data Mining Software: An Update**. *SIGKDD Explorations* 2009, **11**(1):10-18.

20. Frankel AD, Pabo CO: **Cellular uptake of the tat protein from human immunodeficiency virus**. *Cell* 1988, **55**(6):1189-1193.

21. Derossi D, Joliot AH, Chassaing G, Prochiantz A: **The third helix of the Antennapedia homeodomain translocates through biological membranes**. *J Biol Chem* 1994, **269**(14):10444-10450.

22. Morris MC, Depollier J, Mery J, Heitz F, Divita G: **A peptide carrier for the delivery of biologically active proteins into mammalian cells**. *Nat Biotechnol* 2001, **19**(12):1173-1176.

23. Schumacher D, Tischer BK, Teifke JP, Wink K, Osterrieder N: **Generation of a permanent cell line that supports efficient growth of Marek's disease virus (MDV) by constitutive expression of MDV glycoprotein E**. *J Gen Virol* 2002, **83**(Pt 8):1987-1992.

24. Sanders WS, Bridges SM, McCarthy FM, Nanduri B, Burgess SC: **Prediction of peptides observable by mass spectrometry applied at the experimental set level**. *BMC Bioinformatics* 2007, **8 Suppl 7**(8):S23.

25. Kawashima S, Ogata H, Kanehisa M: **AAindex: Amino Acid Index Database**. *Nucleic Acids Res* 1999, **27**(1):368-369.

26. Gulaboski R, Scholz F: **Lipophilicity of Peptide Anions: An Experimental Data Set for Lipophilicity Calculations**. *Journal of Physical Chemistry B* 2003, **107**:5650-5657.

27. Mitaku S, Hirokawa T, Tsuji T: **Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces**. *Bioinformatics* 2002, **18**(4):608-616.

28. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A: **GOR V server for protein secondary structure prediction**. *Bioinformatics* 2005, **21**(11):2787-2788.

TABLE 3.1

CONFUSION MATRICES FOR DATASETS GENERATED USING DIFFERENT APPROACHES.

Dataset 1. Unbalanced (total examples 145).

Non-CPP	CPP	← Classified as
0	34	Non-CPP
1	110	CPP

Dataset 2. Balanced with random peptides as negatives.

A. 10-fold cross-validation with training data (total examples 222).

Non-CPP	CPP	← Classified as
109	2	Non-CPP
7	104	CPP

B. Tested on unbalanced data (total examples 145).

Non-CPP	CPP	← Classified as
12	22	Non-CPP
6	105	CPP

TABLE 3.1 continued

Dataset 3. Balanced with biological peptides as negatives.

A. 10-fold cross-validation with training data (total examples 222).

Non-CPP	CPP	←Classified as
108	3	Non-CPP
10	101	CPP

B. Tested on unbalanced data (total examples 145).

Non-CPP	CPP	← Classified as
10	24	Non-CPP
6	105	CPP

Dataset 4. Balanced by sampling known negatives.

A. 10-fold cross-validation with training data (total examples 222).

Non-CPP	CPP	← Classified as
96	15	Non-CPP
10	101	CPP

B. Tested on unbalanced data (total examples 145).

Non-CPP	CPP	← Classified as
29	5	Non-CPP
7	104	CPP

TABLE 3.2

CLASSIFIER PERFORMANCE WITH DIFFERENT TRAINING REGIMES.

a. Performance from ten-fold cross validation with training data sets.

	Unbalanced	Balanced with random negatives	Balanced with biological negatives	Balanced by sampling from known negatives	Balanced by sampling from known positives*
Accuracy	75.86%	95.94%	94.14%	88.73%	78.82%
True Positive Rate	0.759	0.959	0.941	0.887	0.7883
False Positive Rate	0.768	0.041	0.059	0.113	0.2117
ROC	0.495	0.959	0.941	0.887	0.7883

*- These values represent the averages for 10 datasets. .

b. Performance of each classifier with original dataset.

	Unbalanced	Balanced with random negatives	Balanced with biological negatives	Balanced by sampling from known negatives
Accuracy	75.86%	80.69%	79.31%	91.70%
True Positive Rate	0.759	0.807	0.793	0.917
False Positive Rate	0.768	0.508	0.553	0.127
ROC	0.495	0.649	0.620	0.895

TABLE 3.3

COMPARISON OF SVM BASED CPP CLASSIFIERS TO PREVIOUSLY PUBLISHED METHODS.

	Hällbrink-2005 [7]	Hansen-2008 [8]	Dobchev-2010 [9]	Unbalanced	Distribution-based	Biologically-based	Balanced by sampling Non-CPPs
Overall Accuracy	77.27%	67.44%	83.16%	75.86%	80.69%	79.31%	91.72%
CPP Accuracy	88.46%	80.30%	92.21%	99.10%	94.59%	94.59%	93.69%
Non-CPP Accuracy	35.71%	25.00%	54.17%	0.00%	35.29%	29.41%	85.29%

TABLE 3.4

FEATURES SELECTED FOR DATASETS GENERATED USING APPROACHES 1-4.

Dataset 1 (Balanced with random negative examples)	Dataset 2 (Balanced with biological peptides assumed to be negative)	Dataset 3 (Unbalanced dataset)	Dataset 4 (Balanced by random sampling of known negatives with replacement)
Net Charge	Net Charge	Net Charge	Negative Charge
Positive Charge	Isoelectric Point	Positive Charge	Isoelectric Point
Number of serines (S)	Molecular Weight	Number of alanines (A)	Number of glycines (G)
Number of aspartates (D)	Hydropathicity	Number of arginines (R)	Number of alanines (A)
Percent valine (V)	Number of valines (V)	Percent arginines (R)	Number of tryptophans (W)
Percent proline (P)	Number of lysines (K)	Net Donated Hydrogen Bonds	Number of asparagines (N)
Percent phenylalanine (F)	Number of arginines (R)		Number of lysines (K)
Percent threonine (T)	Percent glycine (G)		Number of histidines (H)
Percent asparagine (N)	Percent methionine (M)		Number of aspartates (D)
Percent tyrosine (Y)	Percent tyrosine (Y)		Percent phenylalanine (F)
Percent cysteine (C)	Percent cysteine (C)		Percent tryptophan (W)
Percent arginine (R)	Percent aspartate (D)		Percent arginine (R)
Percent histidine (H)	Percent negative		Percent histidine (H)
Percent aspartate (D)	Water Octanol Partition Coefficient		Percent Hydrophobic
Percent negative	Net Donated Hydrogen Bonds		Percent negative
Steric Bulk	Percent Helix		Hydrophobicity
Net Donated Hydrogen Bonds	Percent Coil		Water Octanol Partition Coefficient
Percent Helix			
Percent Coil			

TABLE 3.5

FEATURES SELECTED FOR TEN DATASETS GENERATED USING APPROACH 5 – BALANCED SUBSETS OF CPPS SAMPLED WITH REPLACEMENT COMBINED WITH KNOWN-CPP ANALOGS.

Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10
Number (V)	Length	Number (R)	Net Charge	Net Charge	Percent (T)	Net Charge	Positive Charge	Number (W)	Positive Charge
Percent (R)	Net Charge	Percent (W)	Negative Charge	Percent (I)	Percent (Y)	Positive Charge	Number (G)	Number (T)	Percent (I)
	Number (V)	Percent positive	Number (I)	Hydrophobicity	Net Donated Hydrogen Bonds	Percent (I)	Number (S)	Number (R)	Amphipacity
	Number (C)	Amphipacity	Number (H)	Net Donated Hydrogen Bonds	Percent Sheet	Percent (W)	Percent (F)	Percent (S)	
	Percent (H)	Percent Helix	Percent (F)			Percent Hydrophobic	Percent (R)	Percent (T)	
	Net Donated Hydrogen Bonds		Net Donated Hydrogen Bonds				Percent (H)		
							Amphipacity		

TABLE 3.6

KNOWN CELL-PENETRATING PEPTIDES FROM THE LITERATURE AND COMMERCIAL VENDORS.

Cell-penetrating peptide	Reference
AAVALLPAVLLALLAKNNLKDCGLF	[12]
AAVALLPAVLLALLAKNNLKECGLY	[12]
AAVALLPAVLLALLAPVQRKQKLMP	[12]
AAVALLPAVLLALLAVTDQLGEDFFAVDLEAFLQEFGLLPEKE	[12]
AAVLLPVLLAAP	[8, 12]
AGYLLGKINLKALAALAKKIL	[7, 8]
AGYLLGKLKALAALAKKIL	[8]
AHALCLTERQIKIWFQNRRMKWKKEN	[8]
AHALCPPERQIKIWFQNRRMKWKKEN	[8]
ALWKTLLKKVLKA	[7]
AYALCLTERQIKIWFANRRMKWKKEN	[8]
CGPGSDDEAAADAQHAAPPKKKRKVGY	[8]
CNGRC	[12]
CNGRCG	[12]
CNGRCGGKKLKLLKLL	[12]
CNGRCGGKLAKLAKLAKLAK	[12]
CNGRCGGLVTT	[12]
GAARVTSWLGRQLRIAGKRLEGRSK	[7]
GALFLGFLGAAGSTMGAWSQPKSKRKV	[12]
GGRQIKIWFQNRRMKWKK	[7]
GIGKFLHSAKKWGKAFVGQIMNC	[12]
GLAFLGFLGAAGSTMGAWSQPKSKRKV	[8]
GRKKRRQ	[7]
GRKKRRQRRPPQC	[8]
GRKKRRQRRRC	[7, 8]
GRKKRRQRRRPPC	[7, 8]
GRKKRRQRRRPQ	[7, 8]
GRQLRIAGKRLEGRSK	[7]
GWTLNPAGYLLGKINLKALAALAKKIL	[7, 8]
GWTLNPPGYLLGKINLKALAALAKKIL	[7, 8]
GWTLNSAGYLLGKINLKALAALAKKIL	[7, 8, 12]
GWTLNSAGYLLGKINLKALAALAKKLL	[7, 8]
GWTLNSAGYLLGKLKALAALAKKIL	[7, 8]
GWTLNSKINLKALAALAKKIL	[8]
INLKALAALAKKIL	[12]
IWFQNRRMKWKK	[8]
KALAALLKKWAKLLAALK	[8]
KALAKALAKLWKALAKAA	[7, 8]
KALKKLLAKWAAAKALL	[7, 8]
KCRKKKRRQRRRKKLSECLKRIGDELDS	[7]
KCRKKKRRQRRRKKPVVHLTLRQAGDDFSR	[7]
KETWWETWWTEWSQPKKKRKV	[12]
KETWWETWWTEWSQPKKRKV	[8]
KFHTFPQTAIGVGAP	[8]
KITLKLAIKAWKLALKAA	[7, 8]
KIWFQNRRMKWKK	[8]
KLAAALLKKWKKLAAALL	[7, 8]
KLALKALKALKAALKLA	[7, 8]

TABLE 3.6 continued

KLALKLALKALKAALK	[7, 8]
KLALKLALKALQAALQLA	[8]
KLALKLALKAWKAALKLA	[7, 8]
KLALQLALQALQAALQLA	[8]
KMTRAQRRAAARRNRWTAR	[7]
KRPAATKKAGQAKKKKL	[7]
LGTYTQDFNKFHTFPQTAIGVGAP	[8]
LIRLWSHLIHIWFQNRRLKWKKK	[8]
LKTLATALTKLAKTLTTL	[8]
LKTLTETLKELTKTLTEL	[8]
LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTESC	[8]
LLIILRARIRKQAHAHSK	[7]
LLIILRRPIRKQAHAHSK	[7]
LLIILRRRIRKQAHAHSA	[7]
LLIILRRRIRKQAHAHSK	[7, 8]
LNSAGYLLGKINLKALAALAKKIL	[7, 8]
LNSAGYLLGKLKALAALAKIL	[8]
MANLGYWLLALFVTMWTDVGLCKKRPKP	[8]
MDAQTRRRERRAEKQAQWKAAN	[7, 12]
MGLGLHLLVLAAALQGAKKKRKV	[7]
MPKKKPTPIQLNP	[12]
MVKSKIGSWILVLFVAMWSDVGLCKKRPKP	[8]
MVTVLFRRLRIRRACGPPRVRV	[8]
NAKTRRHERRRKLAIER	[7, 12]
PKKKRKV	[12]
PKKKRKVALWKTLLKKVLKA	[7]
PMLKE	[8]
QLALQLALQALQAALQLA	[8]
RGGRLSSYSRRRFSTSTGR	[8]
RGGRLSYSRRRFSTSTGR	[7]
RGGRLSYSRRRFSTSTGRA	[12]
RKKRRQRRR	[7, 8]
RKSSKPIMEKRRRAR	[7]
RQARRNRRRALWKTLLKKVLKA	[7]
RQGAARVTSWLGRQLRIAGKRLEGR	[7]
RQGAARVTSWLGRQLRIAGKRLEGRSK	[7]
RQIKIWFPNRRMKWKK	[7, 8]
RQIKIWFQNMRRKWKK	[8]
RQIKIWFQNRRMKWKK	[7, 8, 12]
RQIKIWFQNRRMKWKKLRKKKKKH	[7]
RQIRIWFQNRRMRWRR	[8, 12]
RQPKIWFPNRRMPWKK	[8]
RRLSSYSSRRRF	[8]
RRMKWKK	[8]
RRRRRRRRR	[7, 8, 12]
RRWRRWWRRWWRRWRR	[8]
RVIRVWFQNKRCKDKK	[7, 8]
RVTSWLGRQLRIAGKRLEGRSK	[7]
SWLGRQLRIAGKRLEGRSK	[7]
TAKTRYKARRAELIAERR	[7, 12]
TRQARRNRRRWRERQR	[8]

TABLE 3.6 continued

TRRNKRNRIQEQLNRK	[7, 8, 12]
TRSSRAGLQFPVGRVHRLLRK	[12]
TRSSRAGLQWPVGRVHRLLRKGGC	[12]
VPALR	[8]
VPMLK	[8]
VPTLK	[8]
VQAILRRNWNQYKIQ	[7]
VRLPPPVRLPPPVRLPPP	[8]
WFQNRRMKWKK	[8]
YGRKKRRQRRR	[12]
YGRKKRRQRRRGTSSSSDELSWIIELLEK	[7]
YGRKKRRQRRRSVYDFFVWL	[7]

TABLE 3.7

KNOWN NON-PENETRATING CELL-PENETRATING PEPTIDE ANALOGS AND PEPTIDE HORMONES.

Non-cell penetrating peptide	Reference
AGCKNFFWKTFTSC	[7]
AHALCLTERQIKSNRRMKWKKEN	[8]
CYFQNCPRG	[7]
DFDMLRCMLGRVYRPCWQV	[7]
EILLPNNYNAYESYKYPGMFIALSK	[7]
FITKALGISYGRKKRRQC	[8]
FVPIFTHSELQKIREKERNKGQ	[7]
GRKKRRQPPQC	[8]
GWTLNSAGYLLGKFLPLILRKIVTAL	[7, 8]
GWTLNSAGYLLGKINLKAPAALAKKIL	[7, 8]
GWTLNSAGYLLGPHAI	[7]
GWTNLSAGYLLGPPPGFSPFR	[7]
HDEFERHAEGTFTSDVSSYLEGQAAKEFIAWLVKGR	[7]
IAARIKLRSRQHIKLRHL	[8]
ILRRRIRKQAHAHSK	[8]
KIWFQNRRMK	[8]
KKKQYTSIHHGVVEVD	[7]
KKLSECLKRIGDELDS	[7]
KLALKALKAALKLA	[7, 8]
KLALKLALKALKAA	[8]
LLGKINLKALAALAKKIL	[8]
LLKTTALLKTTALLKTTA	[7, 8]
LLKTTELLKTTELLKTTE	[7, 8]
LNSAGYLLGKALAALAKKIL	[7, 8]
LNSAGYLLGKLKALAALAK	[7, 8]
LRKKKKKH	[7]
PVVHLTLRQAGDDFSR	[7]
QNLGNQWAVGHLM	[7]
RPPGFSPFR	[7]
RQIKIFFQNRRMKFKK	[7, 8]
RQIKIWFQNRRM	[8]
RQIKIWFQNRRMKWK	[8]
TERQIKIWFQNRRMK	[8]
WSYGLRPG	[7]

TABLE 3.8

A LIST OF INITIAL FEATURES USED FOR CLASSIFIER CONSTRUCTION.

Feature	Reference
Length of peptide	[24]
Net charge of peptide	[24]
Positive charge	[24]
Negative charge	[24]
Isoelectric point (pI)	[24]
Molecular weight	[24]
Hydropathicity	[25]
Number of Each Amino Acid (20 features)	[24]
Percent composition of each amino acid (20 features)	[24]
Percent polar amino acids	[24]
Percent positive amino acids	[24]
Percent negative amino acids	[24]
Percent hydrophobic amino acids	[24]
Hydrophobicity	[25]
Lipophilicity	[26]
Amphiphilicity	[27]
Water-Octanol Partition Coefficient	[25]
Steric Bulk	[25]
Side chain bulk	[8]
Net donated hydrogen bonds	[8]
Percent α helix	[28]
Percent random coil	[28]
Percent β sheet	[28]

TABLE 3.9

PEPTIDES SYNTHESIZED FOR EXPERIMENTAL VALIDATION OF CLASSIFIER.

Name	Role	Sequence (N to C)
HIV-TAT [20]	Control(+)	YGRKKRRQRRR-NH_2
Antennapedia [21]	Control(+)	RQIKIWFQNRRMKWKK-NH_2
Pep-1 [22]	Control(+)	KETWWETWWTEWSQPKKKRKV NH_2
negative-1	Control(-)	TCSSNCQTCPCSSNNCQ-NH_2
negative-2*	Control(-)	GLALLGIAVAILVVL-NH_2
negative-3	Control(-)	PGNIQMMSVVSMSMTITN-NH_2
peptide-1	Predicted CPP	FKIYDKKVRTRVVKH-NH_2
peptide-2	Predicted CPP	RASKRDGSWVKKLHRILE-NH_2
peptide-3	Predicted CPP	KGTYKKKLMRIPLKGT-NH_2
peptide-4	Predicted CPP	LYKKGPAKKGRPPLRGWFH-NH_2
peptide-5	Predicted Non-CPP	FFSLPPVTQDWNSD-NH_2
peptide-6	Predicted Non-CPP	HSPIIPLGTRFVCHGVT-NH_2
TP13 [8, 16]	Known Non-CPP-CPP Analog	LNSAGYLLGKALAALAKKIL-NH_2

Footnote: *negative-2 was unable to be synthesized to desired purity levels due to insolubility issues.

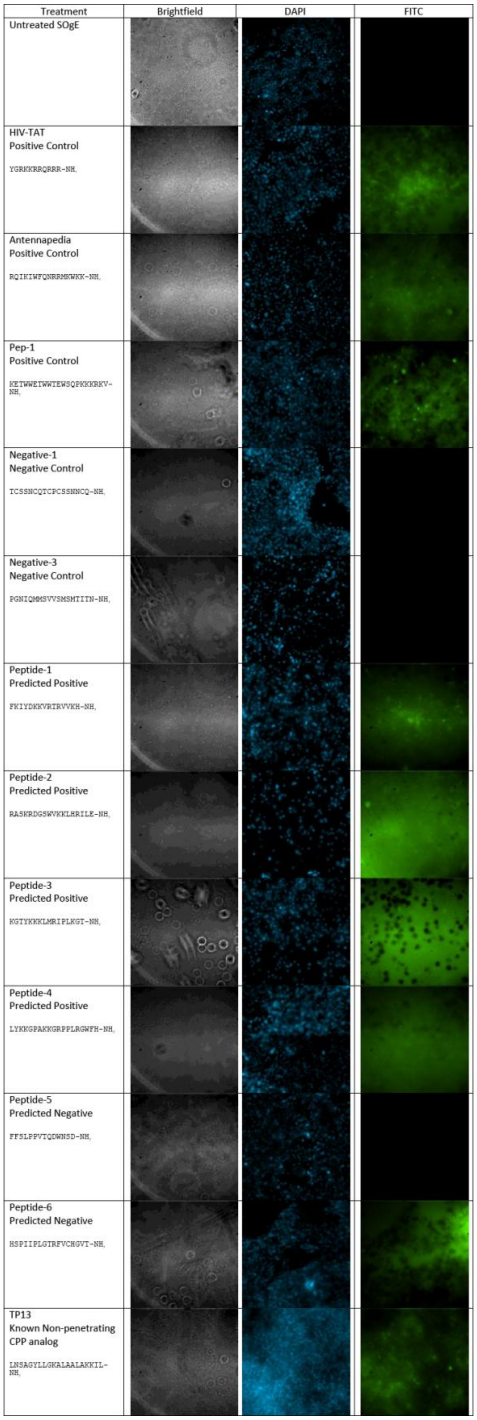


Figure 3.1

Cellular Internalization Microscopy Array of FITC-Labeled Peptides

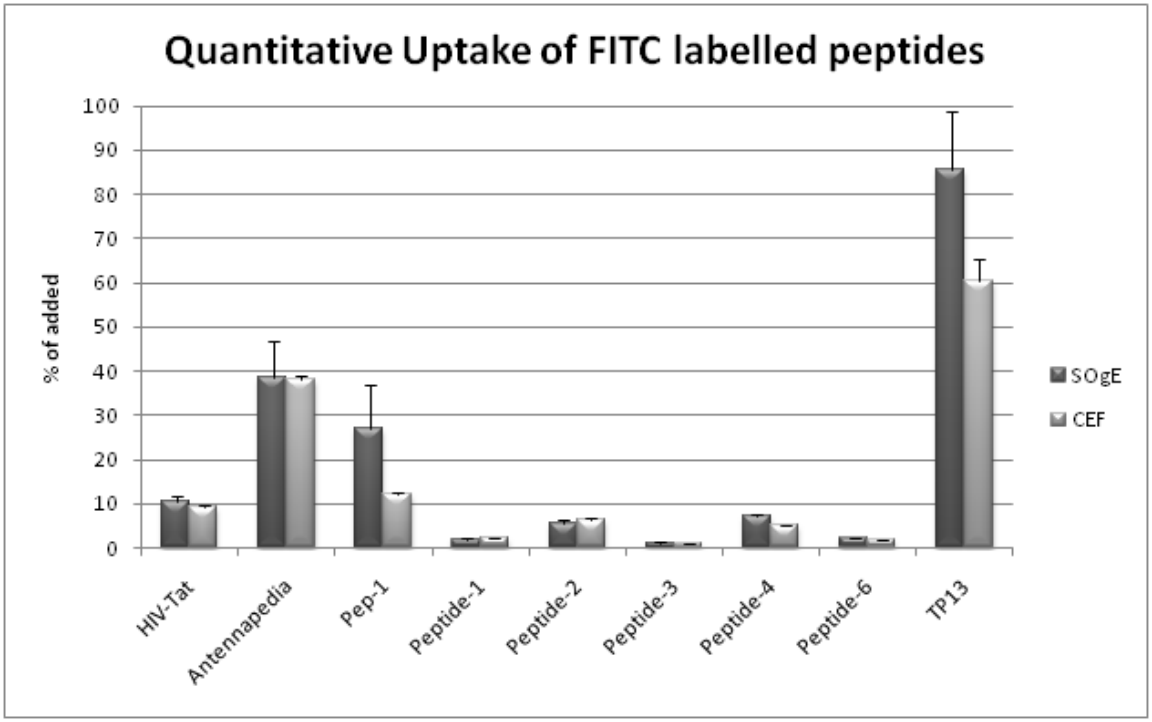


Figure 3.2

Quantitative Uptake Analysis

CHAPTER IV

THE PROTEOGENOMIC MAPPING TOOL

Abstract

Background

High-throughput mass spectrometry (MS) proteomics data is increasingly being used to complement traditional structural genome annotation methods. To keep pace with the high speed of experimental data generation and to aid in structural genome annotation, experimentally observed peptides need to be mapped back to their source genome location quickly and exactly. Previously, the tools to do this have been limited to custom scripts designed by individual research groups to analyze their own data, are generally not widely available, and do not scale well with large eukaryotic genomes.

Results

The Proteogenomic Mapping Tool includes a Java implementation of the Aho-Corasick string searching algorithm which takes as input standardized file types and rapidly searches experimentally observed peptides against a given genome translated in all 6 reading frames for exact matches. The Java implementation allows the application

to scale well with larger eukaryotic genomes while providing cross-platform compatibility.

Conclusions

The Proteogenomic Mapping Tool provides a standalone application for mapping peptides back to their source genome on a number of operating system platforms with standard desktop computer hardware. Researchers are provided with the options for selecting different genetic codes and selecting different methods for determining splice sites. The program executes very rapidly across a wide range of datasets and enables researchers to structurally annotate genomes using MS derived proteomics data in standard format.

Background

Expressed proteins provide experimental evidence that genes in the genome are being transcribed and translated to produce a protein product. Recently, a new structural genome annotation method, proteogenomic mapping, has been developed that uses identified peptides from experimentally derived proteomics data to identify functional elements in genomes and to improve genome annotation [1-2]. Initially used for the structural annotation of prokaryotic genomes, proteogenomic mapping is rapidly gaining traction in eukaryotic genome annotation projects with larger genomes as a complementary method [3-4].

Proteogenomic mapping can identify potential new genes or corrections to the boundaries of predicted genes by using peptide matches against the genome that do not match against the predicted proteome to generate expressed Protein Sequence Tags (ePSTs) [2]. When aligned with the genome and combined with the published structural annotation, these ePSTs are indicative of translation throughout the genome and can serve to supplement traditional structural genome annotation methods [3-5].

While a number of research groups are becoming increasingly active in the field of proteogenomic mapping [1-5], there is a lack of published and standardized tools to rapidly and exactly map identified peptides back to the genome translated in all 6 reading frames. To our knowledge, there is only one comparable tool, PepLine [6], for proteogenomic mapping. PepLine utilizes de novo based spectral identification based on short spectral match translations of 3-4 amino acids with flanking masses on either end for searches against the genome. In contrast, our tool enables the researcher to use the same database search algorithm and peptide validation approach for both protein identification and improved genome structural annotation.

Implementation

The Proteogenomic Mapping Tool is free to obtain and use, is written completely in Java, and is available for all common computer platforms. It is licensed under GNU GPLv3 making the source code available to the end user [7]. We provide both a command line version and a graphical user interface (GUI) for all common platforms.

Data Input and Customization

The GUI (Figure 1) takes three files as input from the user: a FASTA file of the peptides to be searched, a FASTA file containing the nucleic acid sequences the peptides are to be mapped against (typically the genome), and a file containing the genetic code to use based on the format of the National Center for Biotechnology Information's (NCBI) toolkit for genetic codes [8]. Furthermore, FASTA output from the splice site prediction tool GeneSplicer [9] can optionally be provided. If present, the splice sites given in that file are used instead of the default splice sites for generation of ePSTs. The user is also required to provide a file name and location for the three output files that will be generated.

To generate the FASTA file of the peptides to be searched, it is expected that the user will have performed spectral matching of their MS dataset against databases generated from both the proteome and the genome translated in all six reading frames and confirmed these peptide identifications using a peptide validation strategy. After validation, the unique peptide identifications resulting from a database search against the genome that are not contained among the proteome peptide identifications should be used as the list of peptides to be searched.

The command line version of the Proteogenomic Mapping Tool allows the same inputs as the GUI to be specified as command line arguments and can be run on standard computer platforms (Windows, Linux, Unix, MacOS). An example of using the command line version of the program is included in the README file provided with the application.

The application translates the nucleotide database to protein in all 6 reading frames using the genetic code selected by the user. We provide the most common genetic codes from NCBI [8] which are represented in NCBI's standard format for genetic codes in the *genetic_code_table* file included with the application. The tool maps the peptides to the translated genome using the Aho-Corasick string searching algorithm to provide rapid and exact matches of peptides to the genome [10-11]. The Aho-Corasick string matching algorithm [10] quickly locates all occurrences of keywords within a text string. The algorithm consists primarily of two phases. In the first, a finite state machine is constructed from the set of keywords. The time to construct this machine and its memory requirements are linearly proportional to the sum of the lengths of the keywords. The second phase consists of running the state machine using the text string as input. This phase takes time linearly proportional to the length of the text string. Thus, the time to run the entire algorithm is proportional to the sum of the length of the keywords and the length of the text string. In our case, the peptides for which to search are the keywords, and the reference genome against which to search is the text string.

ePST Generation

Once a peptide has been mapped to a nucleotide sequence, the reverse translated peptide is used to create an expressed Protein Sequence Tag (ePST) [2]. Figure 2 illustrates the ePST generation process for prokaryotes and Figure 3 shows both options for the ePST generation process in eukaryotes. For prokaryotes, the reverse translated peptide is extended in the 3' direction to an in-frame stop codon. In the 5' direction, the first in-frame stop-codon upstream of the peptide (5' stop) is identified and the peptide is

extended to the first in-frame start downstream from this 5' stop before the start of the peptide. In the case that no in-frame start occurs between the 5' stop and the start of the peptide, the start of the peptide is used as the start of the ePST. The process is more complex for eukaryotes due to splicing. For eukaryotes, the peptides can be extended to produce ePSTs using three different approaches. In the first approach, the peptide is extended downstream to the first in-frame stop or splice site signal [12] and upstream until the first in-frame start, in-frame stop, or splice site signal. We have found that this approach often generates ePSTs that are far longer than typical exons. We speculate that this is because the potential new ORFs identified by this approach do not have a canonical splice site signal. While the application does default to using canonical splice site signals, our second approach includes the option of using predictions from GeneSplicer [9], a computational splice site prediction tool. The user can select to input GeneSplicer output for use instead of the canonical splice site signals. A third option is to extend the peptide upstream and downstream by a nucleotide length specified by the user given as the number of codons.

Output File Description

Three output files are produced by the application. The first file is a FASTA file containing the ePSTs generated for the dataset. The second file is a more detailed tab separated text file containing the original peptide identifier from the FASTA header, the peptide sequence, the FASTA header for the nucleotide sequence containing the match, the mapping start and end locations for the reverse translated peptide, the strand of the nucleotide match, the reading frame of the match, the reverse translated peptide

sequence, a longer nucleotide sequence extending from the 5' in-frame stop codon immediately upstream of the peptide to the 3' in-frame stop codon immediately downstream of the peptide, the ePST nucleotide sequence and the start and stop locations of the ePST on the nucleotide sequence, the length of the ePST, and the translated ePST. The third file is a GFF3 file containing the ePSTs generated for the dataset to provide researchers with a file format they can quickly load into genome browsers for data visualization.

Example Datasets

To test our implementation we acquired previously published proteogenomic mapping datasets for a number of organisms. For a relatively small example data set, we selected a proteogenomic mapping dataset for the channel catfish virus [5]. This small dataset contains 407 unique peptide identifications, of which 17 peptides did not map to the predicted proteome of the virus, but do map to novel open reading frames in the viral genome. The expression of several of these genes was confirmed by RT-PCR [5]. Our example dataset consists of a FASTA file of these 17 peptides and the reference genome (NC_001493.1) for the channel catfish virus. For bacterial examples, proteomics datasets from three different microorganisms [13] were used to test our application: *Histophilus somni* strain 2236, *Mannheimia haemolytica* strain PHL213, and *Pasteurella multocida* strain 3480. For a eukaryotic example, a previously published proteomics dataset generated from chicken serum was utilized for testing [14]. Table 1 details the number of unique peptides and the number of unique peptides mapping uniquely to the genomic database search contained in each of these five datasets.

Results and Discussion

The output from the Proteogenomic Mapping Tool matches the previously published results against the CCV test dataset [5], and our output provides additional information that not only places the mapped peptides on the appropriate nucleotide strand but also includes the reading frame in which the match occurs. Table 2 gives a list of the peptides and corresponding ePSTs for this dataset. We have also successfully tested this tool for proteogenomic mapping in previously published bacterial [13] and eukaryotic datasets [2, 14]. Table 3 provides runtime analysis for each of our five test datasets, and demonstrates that the Proteogenomic Mapping Tool scales well for increasingly large datasets.

Possible future updates to this application include parallelization of the searches against the genome in all 6 reading frames, and the introduction of better thread support to improve performance further on today's modern increasingly multi-core processors.

Conclusions

The Proteogenomic Mapping Tool is a standalone program that facilitates a streamlined mapping of peptides to a target genome for structural genome annotation through the use of proteomics. This software can be used on a variety of current operating systems and is its ability to use a variety of genetic codes makes it easily customizable for researchers performing proteogenomic mapping in a variety of prokaryotes, eukaryotes, and viruses.

Availability and Requirements

Project name: The Proteogenomic Mapping Tool
Project home page: http://www.agbase.msstate.edu/tools/pgm/
Operating system(s): Windows XP, Vista (x86), Vista(x64), Linux, MacOS
Programming languages: Java
Other requirements: Java
License: GNU GPLv3 [7]
Any restrictions to use by non-academics: None

LITERATURE CITED

1. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation.** *Proteomics* 2004, **4:**59-77.

2. McCarthy FM, Cooksey AM, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a whole organ using proteomics: the avian bursa of Fabricius.** *Proteomics* 2006, **6:**2759-2771.

3. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics.** *Proc Natl Acad Sci U S A* 2008, **105:**21034-21038.

4. Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, Stephenson JL, Jr.: **Whole genome searching with shotgun proteomic data: applications for genome annotation.** *J Proteome Res* 2008, **7:**80-88.

5. Kunec D, Nanduri B, Burgess SC: **Experimental annotation of channel catfish virus by probabilistic proteogenomic mapping.** *Proteomics* 2009, **9:**2634-2647.

6. Ferro M, Tardif M, Reguer E, Cahuzac R, Bruley C, Vermat T, Nugues E, Vigouroux M, Vandenbrouck Y, Garin J, Viari A: **PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences.** *J Proteome Res* 2008, **7:**1873-1883.

7. **The GNU General Public License version 3** [http://www.gnu.org/copyleft/gpl.html]

8. **NCBI Genetic Code Table** [ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt]

9. Pertea M, Lin X, Salzberg SL: **GeneSplicer: a new computational method for splice site prediction.** *Nucleic Acids Res* 2001, **29:**1185-1190.

10. Aho AV, Corasick MJ: **Efficient String Matching: An Aid to Biblographic Search.** *Communications of the ACM* 1975, **18:**333-340.

11. Dandass YS, Burgess SC, Lawrence M, Bridges SM: **Accelerating string set matching in FPGA hardware for bioinformatics research.** *BMC Bioinformatics* 2008, **9:**197.

12. Wu Q, Krainer AR: **AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes.** *Mol Cell Biol* 1999, **19:**3225-3236.

13. Nanduri B, Wang N, Lawrence ML, Bridges SM, Burgess SC: **Gene model detection using mass spectrometry.** *Methods Mol Biol*, **604:**137-144.

14. Corzo A, Kidd MT, Koter MD, Burgess SC: **Assessment of dietary amino acid scarcity on growth and blood plasma proteome status of broiler chickens.** *Poult Sci* 2005, **84:**419-425.

TABLE 4.1

EXAMPLE DATASET STATISTICS.

	Channel catfish virus	*H. somnus* 2236	*M. haemolytica* PHL213	*P. multocida* 3480	*G. gallus* serum
Number of unique peptides	407	958	1,755	675	1,447
Number of unique peptides mapping exclusively to genome	17	305	1,585	376	92

TABLE 4.2

CHANNEL CATFISH VIRUS PEPTIDES AND ePSTS.

ID	Peptide	Reading Frame	ePST
Proteinase-1	NLDLLDNSTG	+1	CTGCTGACCAGGCTACTGTTTGTATGCACAAT CTTGACCTTCTCGACAATTCCACTGGTGCTCCA CAAGGGGATCTCACCGATCCAAGAGAAGATG GGTAGG
Proteinase-2	LMPCSMSS	+1	ATGATCCGGACGAGGTTCCTAGTTCGAAGAGA GGGCCTTCTCGATGTGGTCTCTCCCGGTGAACT CTTCTCCGGAGAACACGGGGTAATCACCCCCG GGACTGAACGATATAGACTCATGCCATGTTCC ATGTCCTCTATTTGAT
Proteinase-3	PSPVSSHPLAASVSGPC	-1	GTGATTCTTCGTCTTTCCGAGCCCCGTATCGTC GCACCCATTAGCCGCTTCGGTGAGTGGACCTT GTGTCGCAGACATCTTCAAGACAAGCGATTGG TTCAGATGGTGGAATTGGAATGAATATTCGCG TATATTCACCAGTGTCTTTTAAT
Proteinase-4	MRELVSM	+3	TTGATGTTTTTGTTCCCGTCTCTATATCTTTATT CAGAGTCTGAACCAGTGACACTTAGATTGTTA TCATATGATTTAAACCATGATAGGTCACCATC TGTAAATTCCTCATGGTTCATGATCCCGTGCTT GGCACATATCATTATCAGAAGGATGGCCTTCA TCGACAGCTCCACTCTCTGGTGGTCTCTGTCAC TCACCGGCGTGCCCGGGGTCGCGTATTCCACC GCCGTGTCTCTGTTCAAGACGGCGAGTTGGCC TCTGGGGATATCGGCCGCCGTGACGGTCAGGG AGTTGATGAGAGAACTGGTCTCCATGTCAGTG TTTAGTCTCTGGAAGATTTCCTCAGCGGACATC TCGGGTCCCGCTGCTAATGCGAGCCTCAGGGT TTCACGGGTAATCGATAGATGCACCCGCTTGT GGCTATGCCGGGCGGCCGGCCTCTTTCCTCGT ACACGCGGGGTTGGTTTGGGTTCGGCCACGTG CGCGCCCCGGCGTTCCAGTAACGTAACCGGAC GCCTCGAGGGGACCCGCGCGGGCTCGGGATCG GCCCCGATACCACCGGCCGGGACACCGATCAG TTCCAGTGGCCCGCCCGCAGACGGTGGGTCTT CGTCCTCGCTCTCTTCGCTCTCCTCCTCGCTCT CCTCCTCCTCGCCTCCACTCTCCGTCTCGCCCC CTTGTCTATCCTCCTCGTCCTCCTCTCGGCACA CTCCATCTCCGCGGGTGCCGTTCGAGTCCGGC ACCGGATCGACACTCTCATCGTCACCCGATTC CTCACTGCTGAGCTCACGACCACCGGCGTACG ATCCGTGGTAGT

TABLE 4.2 continued

Proteinase-5	RNDIAESSCLVA	-1	TTGATGACGTCCCAGTTCGCCAGGTCGGGTCTC ACCATCGAGAGAAACGACATCGCAGAATCCAGC TGTCTGGTCGCGACCATCGACTCCATGGCCTCG GCGAGACTCGTTCCTTGAT
Proteinase-6	ISRDSIPILF	+3	ATGCTGACACACACCACCCCGAACAAGGCTGTA CGTATCAGGTGCATCAAACCCAGGATACTCGCG GGGGGTGTTCCGGGTGTAGCTCTCACTACATAC CGAAATTTTCCGAGGTCGGAGAGGTCGCTGCAG CTGTTGTGCTTGGTGCCGGATTGTGTGGCCCCCT TACCGGTACTGTTGACAGTCAGCGTTCCGAACT CGGTGAATTCGGTACTGTTGTACACAGACCACA GGCAGTTGACAGGGAAGACCTTCCCGGGTTCTC TCTTTTCGGGTATCTCTAGGGATTCAATCCCAAT CTTGTTCAACCACTCGATGAAGGTGGTGGGTCC CTGTTGGTTGTAGA
Proteinase-7	QAVVPMNTF	-2	CTGTGCGTCAGTTGCTGTAACTTGACATCCGGGT TATCGGTTGGTTTCACCGATAGATCGACCGTGA ACGGACCCGGGGGTAAATCGGCGGGCGCGACCT GCAGGGCCGCTCCGCAAGCGGTCGTCCCCATGA ATACGTTCGAGCATATCACCGCCACATGTGCGT CCTCGAGGTAGT
Proteinase-8	QLGDGPLGGGHVDHIPF	+3	ATGTAGATGACCATGTCCAACTTGAGAGGTCCA ATGTCTACCCCCCGTGGGTCGTGGTACAGAATG TGTGTGTTGTACATGTTCGTTATGAAGTTGATTC CATTGTCTCGGAACGCGACGGCGAGCGAGATGA GTTGTTTCAGGATCACGGCCCCCAGGAGGGTTC CATCGTCCGTCTCGCCATCGAAGTTCAGCTCGGT GATGGACCTCTTGGCGGGGGTCACGTAGATCAT ATACCTTTTCGTGCCATGGGCGCCCAGGGTGTA GTGGAACACCGGTACCACATATGGGATAGATTT TTGGATCGCCTGGATCCCCTCCATCAAAGACTGT ATCGCCTCGAAATCAATGGTCGGTTTCAGCTCG AGGTAGACGATGTCATACGTAGGGGAGAATTCG GGGGGCCTGTATACCCTGATCTCCTTCACGCCGT ACTCTTTGGTGGCCACGACCGGGTACACGGAGA CGAGGTCCCGGGGGGTGAGGGTTATCAGTTTCT TCGCGGTGTGATCATCGCCCATGTCTGCGCGCG CAAGCCATGGCATGTATAGC
Proteinase-9	ARDLPRRF	+2	CTGTGAACAAATATATCTTCGAAGTTTGCCGCG AGGGTACCGACGAGGTCCCGCACGCGATCTACC AAGACGGTTTCCAGGACGTGTCTCACGACTGGA AGGGCCGGGCCCGGCCATATCACCACGATCGAA CCCGGGTCGAGCGCGCACTCGATAGA
Proteinase-10	EVVILQ	-1	ATGCGTACACCGCATACGCCTTCAGCACTGCAC TGTCACGGCTCAGGTCCCATTTACGACGTGCGG GGTAAGGCCTGTCTCCCTTCAGAAATTGCGTGA GCTCGTAGTATTCGCTCAGCACCCTCTGTCCCAG GAACTGGCGTATCCGAGGACAACCACCCCTCGA ATGGTACACGTGTTCGTCCAGGAAATCATCGAC GAGCGTGAAGCGGATGACCTTGACACCGCAGTC TGGACACACGTGACGATCGCTCTTCACATCGTC CGGGATCAAACCTCCCTTGGGTCCGAAGTACAG TCTCGTCATGAACACGAGGTTGTCATCCTGCAA GGTACCGTGGGCGACTATTGTATCGTAATCCAA GGTAACATCGCAAAACCACACACCTCCGTCCGC ACGCCATCCGCTTGGCTTGAGCATTCCCCTGGGT GCCGCGGACCATCTGAACCCCCTGTCCGTGGGG TTGCTGACCCGCTCACCGTCTGTCACCAGGGAA CCGCATTCGAAATCCATCGCCTCAGTAGTGGAT TGTCAGAGATCGTTCTATGGGTATCTGGTCAGTG TGAATTATTGGAATGGGCGCTCGCAGTATTCTTC AATCGTTCTTTTTCGGGCACCATGAGACTCTCGG GATCGAGGAAGCCGCCGGCGGTCCACCGGATGC GACACGTGAGATCCGATAACCTATAAA
Trypsine-1	IPFVSGLMNAQIILFSGPCMIGRNAAVSCK	+3	CTGACAGCCACGGAATCATCGGGGGTGTACACA ACTTCCGAATCCACGGAGTCCATCACCGCGGTG GCGATCCCGACCATCGCACGGAGTTCGGCCTCG GTCCCGATCTTCTCAAGGAAAAAACCAAGGTGT TCCGGGTATACCTTTAGAATTCCCTTCGTTTCCG GGTTGATGAACGCGCAGATCATCTTGTTCTCGG GTCCTTGTATGATAGGCAGGAACGCGGCCGTCT CGTGCAAGCTATCGAAACGATCCATATCATGGG CACCGCCGATGAGATCCATCCCGATGTTCTTGC GGAGTGCATCCATTTCGCTCACAAGAAGATAAA

TABLE 4.2 continued

Trypsine-2	ARTVFLNVRPGWSR	+3	GTAGAGGAGGGCCCGAACCGTCTTCCTAAATGT GAGACCGGGGTGGTCTCGGAAGGACGACGCGTC AGTCGGGCAACCCGCCATCGTACCGGCCAAGAG GGACTTGACACAGGTGCGGATCATTCCCACCTT GTATCCGATCTGGATCGCCCCCTGTGTGACCTGG TTGGTCAGGCTCATGATCTGTTGGTACCACTGCT GGTACGCCATCACTTCCCCGAGGCGCTCGTTGG TCGTCGTACCCAGCTCCTTCAGGCCAGTAGCTA AACGCTTGAAGTTTGTATCCATGGCCAGCATCT GGAGGTTTATCTGGTTTTGTAGGTCGTCCACCCT CCCGTTCACTGCCCTGATGTTGCGGTCTAACTTG GCCGATATCAGGGCGATACTGTCTCCCAGTTCC GTGATCACATCCGCGGTCTTGTCCAATTGTGCCT GTAAACCGTCTATTTTGGAGGCTGCCAATGTGG CGACCGCCAGTGCTGCCGTCGACGCGACGAGTG CCGCGCTGGACATGGTTATCGCGGCCACCGATA ACCCGAATTTATCGCTCGTGGGCACCGATCCGC CCGAGGCGGGCGCGAACATCTTAACCTTTTCGT GTTCGAGGTCCAGGTCAACGAGGCTATTTTTCA ATAGTTCGTTACTCCGCTGGAAGTCCAGGAGTA TCGCCGCCGTCTGAT
Trypsine-3	EGQAQRTCAYPSAGLLQASQGR	+3	CTGTGAAGCCGGCCGTGAGGGACAAGCGCAAC GAACATGTGCCTACCCCAGCGCTGGTTTACTTCA AGCATCACAAGGCCGAGCTGGCCAAGGCGCTGG TTGAG
Trypsine-4	PCSRTSGSGACSGR	-1	CTGCGTAAGACGGAGGAGACCGTGCTCGCGGAC GAGCGGTTCCGGGGCCTGCTCGGGCCGGAGATG GTGGCACGGCTATTGAA
Trypsine-5	NRTRVYTMPGWR	-2	TTGGGTATCAGCTTCCGTCCGCCCCCGGAGCCG CACTCGGGACACTCCCGGGGGGTGCAGAAGAA ACAGAACACGTGTTTACACGATGCCCGGTTGGA GGGAAACACGGCCTCCCCCTGGCAGAAACAACA CGGAAAGACTGGAGACATGAT
Trypsine-6	LKSPPGLRK	-1	CTGGAAAGGCTGAAAAGTCCACCGGGACTGCGA AAGTGAC
Trypsine-7	VARGEDATCPNDKGSEPR	+3	CTGGAAACAGAACTTCTCGAGGCCATCCGAGAC GGTGTCGCGGGTGAAGAGTTTCGCGTGGCAGCC CCTCCGCGCGGGAATGACCACGACGCTGCACTC CAGTATCACCTCGTTGAGGCCCACATCGAGGGT TCCGAGGCTCAGTGCACATGGTATGTCCGCTTCC GTGAACGTCTCCACGCATCTCTTGCCGGTGTCCT CGGACTCATCTATCCCTCCTATCATGTTCAGGTA GACTCGGTCTTCCATGTGGACATGCCAGTAACC GAGGACCTTGCCCATGGGGATCTGGTGCGAGTA CTTGAGCGTGCCCGGAGCGACCTGTACCATTGT TTGGTGGGCCTCGATCGGCTGCTGGTACTTGCGC ATGTGCGCGGGGATGGAGGGGTCGTCGACCGGG TCGCCCGGGGAGAAGATGCAACATGTCCCAACG ACAAAGGTTCCGAACCCCGTGAGTGGAACCTCG TATAGC

TABLE 4.3

RUNTIME ANALYSIS FOR EXAMPLE DATASETS.

Dataset	Genome Size	Number unique peptides mapping to genome	Runtime (ms)
CCV	0.1-Mb	17	563
H. somnus 2236	2.3-Mb	305	2,932
M. haemolytica PHL213	2.8-Mb	1,515	4,507
P. multocida 3480	2.5-Mb	201	3,003
Chicken Serum	1,050-Mb	92	127,991

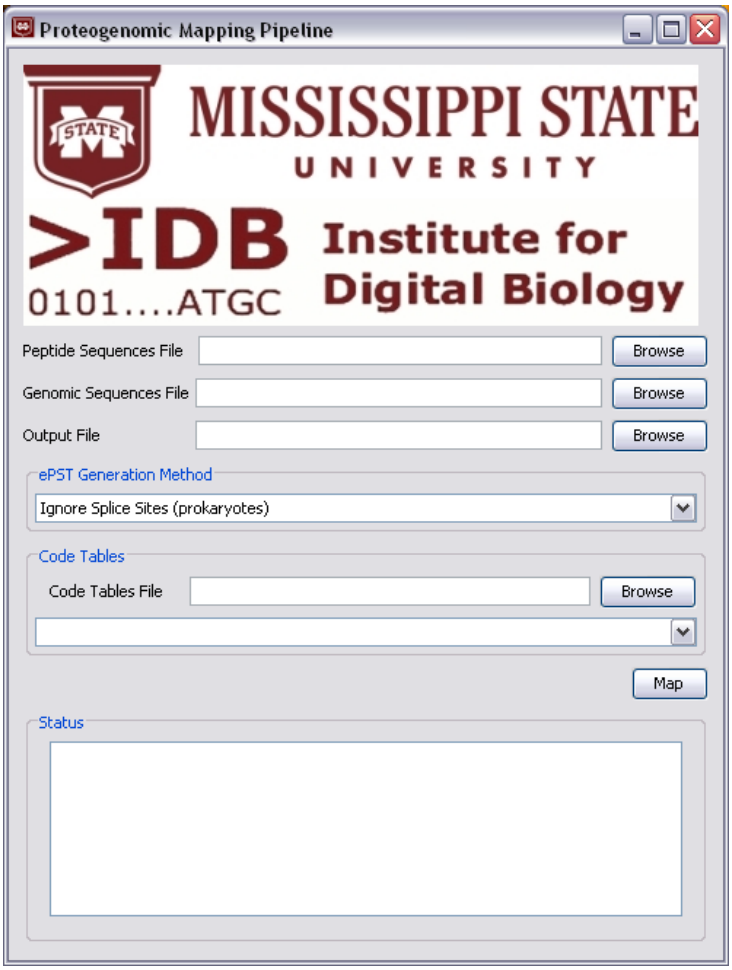


Figure 4.1

Proteogenomic Mapping Tool Windows GUI.

The proteogenomic mapping tool requires the user to provide three files and offers several options:

a. Peptide Sequences File: a fasta formatted file specifying the peptide sequences for which to search.

b. Genomic Sequences File: a fasta formatted file specifying the genome in which to search for the peptides. The file can contain the entire genome as one large entry or multiple entries containing only selected features of interest. For example, the file may contain all exons for an organism.

c. Output File. Two files will be created. The filename provided by the user will contain detailed information about the mapping. An additional fasta file with ".fasta" appended to the name provided by the user will contain the ePST sequence in fasta format.

d. ePST Generation Process: The user is presented with four choices:

    1. Ignore splice sites (prokaryotes)

    2. Use splice sites (eukaryotes)—uses canonical splice junctions to terminate ePSTs.

    3. Use calculated splice sites (GeneSplicer output)

    4. Fixed distance (number of codons)—generates an amino acid sequence of the specified length in both the upstream and downstream direction.

e. Genetic Code Table File: specifies the mapping from codons to amino acids as well as start and stop codons. The genetic code table from NCBI is provided as the default and will typically be selected unless the user is working with an unusual organism. Once the Code Tables file has been selected, the codon table names appearing in this file will be presented as options and the user should select the appropriate codon table (Standard would be used by most researchers. If the user provides the name of a file a different table in NCBI format, the names of all of the codon tables specified will be listed.

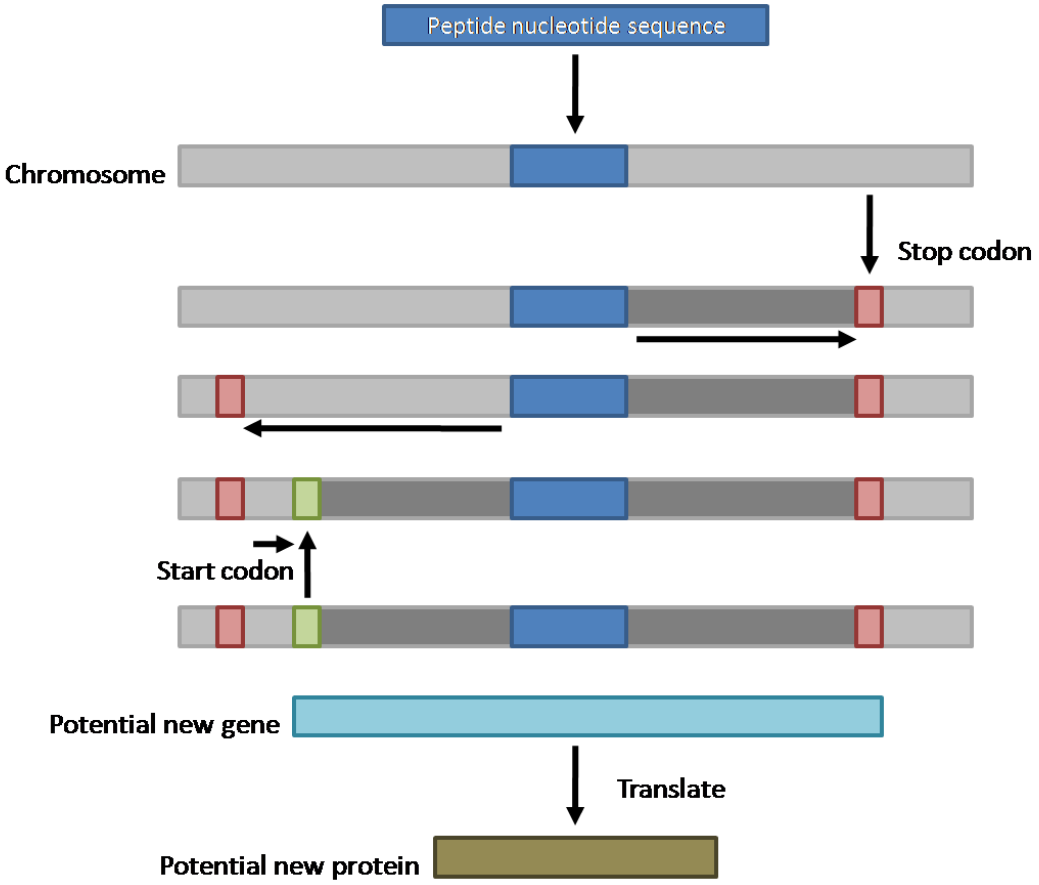


Figure 4.2

Prokaryotic ePST Generation Process.

a. Map the peptide to the translated genome.

b. Extend the mapped peptide in the 3’ direction to an in-frame stop codon.

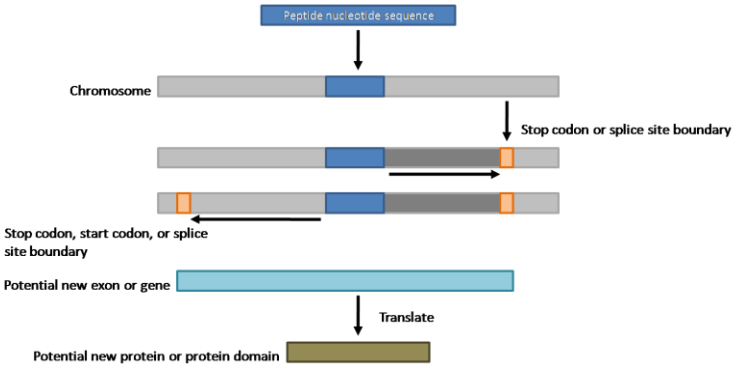
c. Extend the mapped peptide in the 5’ direction to an in-frame stop codon.

d. From this 5’ in-frame stop codon, proceed in a 3’ direction to identify an in-frame start codon.

e. Final ePST.

f. Generate translated ePST sequence.

a)



b)

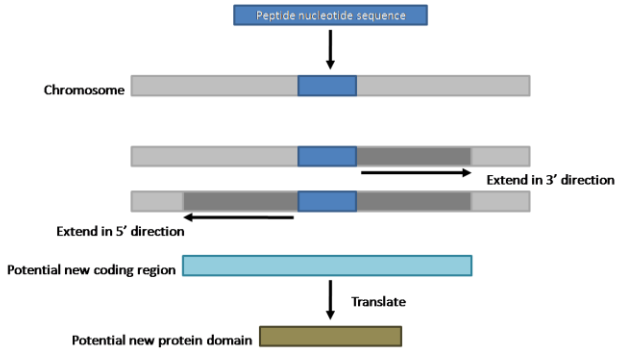


Figure 4.3

Eukaryotic ePST Generation Process.

a. Options 1 and 2: Map the peptide to the translated genome.

b. Option 1: Extend the mapped peptide in the 3' direction to an in-frame stop codon or splice site boundary. Option 2: Extend the mapped peptide in the 3' direction the number of codons selected by the user.

c. Option 1: Extend the mapped peptide in the 5' direction to an in-frame stop codon or start codon, or splice site boundary. Option 2: Extend the mapped peptide in the 5' direction the number of codons selected by the user.

d. Final ePST.

e. Generate translated ePST sequence.

CHAPTER V

PROTEOGENOMIC MAPPING OF *GALLUS GALLUS* SERUM

Abstract

The process of using mass spectrometry derived proteomics data for genome annotation is called proteogenomic mapping. Proteogenomic mapping can make significant contributions to the structural annotation of genomes through the discovery of new functional elements, confirmation of hypothetical and predicted functional elements, corrections to the intron/exon boundaries of known functional elements, and characterization and discovery of alternative splice forms. We use serum proteins derived from *Gallus gallus* (chicken) and mass spectrometry for proteogenomic mapping of expressed peptides to the chicken genome to improve structural annotation. We confirm the expression of 268 proteins from chicken serum and identify an additional 47 peptides that confirm the expression of mRNA, identify novel exons or genes, indicate expression of repeat regions, and correct the boundaries of known exons.

Background

Structural genome annotation is the process of identifying all of the structural elements that comprise an organism's sequenced genome. These structural elements can include regions that code for proteins, both coding and non-coding RNAs, regulatory

regions, and DNA binding motifs. Traditionally, this has been accomplished through the use of expressed sequence tags (ESTs) and cDNA libraries (transcribed RNA that is reverse translated into DNA sequences). These ESTs and cDNAs generally represent approximately 500-800 base pair mRNA sequences that are sequenced as mRNA, or translated back into cDNA and then sequenced [1, 2]. These EST and cDNA libraries are then aligned with the sequenced genome to identify regions representing exons and whole genes that are actively transcribed [1, 2].

These methods are traditionally complemented by the use of computational gene finders that utilize the EST and cDNA libraries and the sequenced genome to identify patterns within the genome indicative of coding regions. This is known as homology based computational annotation [1, 2]. Additionally, these programs can perform *de novo* based genome annotation where they detect signal information within the genome and use these signals to predict coding regions [1]. These computational gene prediction tools produce a number of errors, and significant resources are dedicated to identifying and correcting these errors within the genome annotation [1, 3]. It is estimated that the exact genomic structure is only correctly identified by computational gene finders 50-60% of the time within the human genome, the most well sequenced and annotated genome [3]. These errors can arise from a number of causes. Homology based annotation identifies new genetic sequences based on their similarity to known gene sequences through a combination of similarity information with signal information, and while these methods are very good at identifying new genes similar to known genes, they are limited when given a signal with no similarity information [1]. Several well-known

tools implement a homology based approach to computational gene finding including INFO, ICE, AAT, SYNCOD, EbEST, Est2genome, TAP, PAGAN, DIALIGN [1]. *De novo* based annotation methods typically use signal information identified through the use of Hidden Markov Models (HMMs) to predict genes, and a number of tools utilize this approach including Genscan, Genie, GeneMark.hmm, and FGENESH [1]. Several issues can affect the accuracy of these *de novo* prediction algorithms and give rise to errors including large genes, large introns, highly conserved introns, small exons, overlapping genes, polycistronic gene arrangement, frameshifts, and alternative splice sites [1].

In addition to these traditional structural genome annotation methods, the use of high throughput shotgun proteomics data derived from mass spectrometry experiments is increasingly being used as a complementary method for structural genome annotation [4]. This use of proteomics data to aid in genome annotation was first reported in 2001 [5] for several prokaryotic projects, and was popularized in 2004 by Jaffe et al., who coined the term proteogenomic mapping [6]. Proteomic evidence, identified as expressed Protein Sequence Tags (ePSTs), provides proof that a given gene is expressed, and when back translated and aligned with the sequenced genome, provide structural annotation information for a genome's functional elements [4, 7]. This can include *"confirmation of translation, reading-frame determination, identification of gene and exon boundaries, evidence for post translational processing, identification of splice-forms including alternative splicing, and also, the prediction of completely novel genes"* [4]. Proteogenomic mapping has been utilized in a number of both prokaryotic [5, 6, 8-14]

and eukaryotic [15-26] genome annotation projects, and is increasingly becoming a part of standard annotation pipelines utilizing multiple sources of evidence (sequenced nucleic acids, computational gene prediction, and proteomics data) [2].

Prokaryotic Proteogenomic Mapping

Much prior work in proteogenomic mapping has been done in prokaryotic genome annotation projects [5, 6, 8-14]. These prokaryotic genomes have relatively simple genome structures compared to eukaroytic genomes. Unlike eukaryotes, prokaryotes do not have an intron-exon gene structure nor are they subject to alternative splicing [5]. In addition, these prokaryotic genomes are significantly smaller than eukaryotic genomes, and this small genome size compared to that of eukaryotes allows for direct searching of spectral databases made up of peptides generated from the genome sequence translated in all six reading frames [5]. Table 5.1 shows a comparison of genome sizes for selected prokaryotic and eukaryotic genomes.

Eukaryotic Proteogenomic Mapping

While eukaryotic proteogenomic mapping projects began shortly after their prokaryotic counterparts [15-18, 20, 22, 23], the complications arising from the larger and more complex genome structure has prevented proteogenomic mapping from becoming a significant part of genome structural re-annotation projects in eukaryotes until recently [19, 21, 24-26]. These differences in eukaryotic genome structure compared to prokaryotic genomes arise from the intron-exon structure of genes, repetitive

regions of the genome, gene duplications, splicing and alternative splicing events, and the large areas of intergenic DNA [5].

A number of different approaches for constructing and searching databases of peptide spectra have been developed to address the challenges of using proteomics for genome annotation in eukaryotes. One of the simplest approaches is to search spectra against the genome in its entirety or against selected chromosomes in the genome of interest [15]. A modification of this method is to break the genomes into large chunks of nucleotides with each chunk having some overlap regions with the adjacent chunks [17, 20, 23]. In 2005, Kalume identified 50 novel transcripts and one novel gene in *Anopheles gambiae* (mosquito) using this method [23], while McCarthy (2006) used this method to identify 521 potential novel proteins from the *Gallus gallus* (chicken) "unassigned chromosome". The "unassigned chromosome" represents 10-11% of the chicken genome and is composed of sequences not mapped to the genome assembly [22]. In 2006, Fermin generated a database composed of all ORFs from the *Homo sapiens* genome, and utilized that for DB searches, identifying 282 significant ORFs with 627 novel peptides [18]. In 2007, Tanner utilized computational gene prediction software to identify exons within the *Homo sapien* genome and then constructed exon-splice graphs, which for a given gene take the starting exon and construct sequences by mapping it to all possible internal exons in order to represent alternative splice forms [21]. Using this method, they identified 16 novel genes and extended exons, while confirming over 40 alternate splicing events [21]. Some eukaryotic projects have used the *de novo* sequencing of peptides instead of a database search, and then used these *de novo* identifications to

search genome sequences with BLAST in order to identify regions that code for proteins [16].

More recently, several of these approaches have been combined or coupled with changes to spectral generation protocols by various research groups for more comprehensive proteogenomic mappings [19, 24, 25]. In an application of proteogenomic mapping to the *Arabidopsis thaliana* genome, Castellana (2008) constructed three separate databases: the proteome database, a database comprised of exon-splice graphs, and the genome translated in all six reading frames. Using this multi-database approach, they identified 778 new protein coding genes and refined the annotation of 695 gene models [19]. In a proteogenomic mapping project with *Caenorhabditis elegans* , Merrihew (2008) constructed databases for the proteome, predicted genes from a computational gene finder (GeneFinder) not contained within the proteome database, and the intergenic regions that shared a high homology with *Caenorhabditis briggsae*, a closely related species [25]. Searching against these databases, they identified 429 new coding sequences not present in the known proteome, 33 of which were predicted pseudogenes and 245 of which were novel genes [25]. In 2008, Sevinsky combined isoelectric focusing of peptides subjected to mass spectrometry with databases constructed from a six frame translation of each contig of the *Homo sapiens* genome. These databases were *in silico* trypsin digested and the *in silico* peptides were sorted by molecular weight (MW) and isoelectric point (pI) [24]. These were further separated by splitting the *in silico* peptides into separate databases for every 0.01 pI interval. These were further separated into genic and intergenic databases [24].

Experimental spectra from a given pI range were then searched against the corresponding database, and this methodology yielded 540 genome specific peptides that had no matches against the human proteome [24].

Gallus gallus *Proteogenomic Mapping*

The *Gallus gallus* (chicken) genome draft sequence was released in 2004, and is approximately 1,200 Mbp with ~20,000 to 22,000 genes [27]. The most current build is Build 2.1, released in November 2006, and has a 6.6X coverage with 95% of the genome anchored to chromosomes. The chicken genome contains 38 pairs of autosomal chromosomes and 2 sex chromosomes (Z and W) [27]. Of the 38 autosomal chromosomes, 33 are classified as microchromosomes and these microchromosomes have a very high gene density [28]. A large portion of the unsequenced genome resides on the microchromosomes and this results in ~5-10% of the predicted chicken genes being absent from the Ensembl gene set [27]. Chicken represents an important agricultural species, has a long history as an important medical model, serves as the avian model organism, and it is an important vertebrate outlier on phylogenetic trees because of its evolutionary distance (~310 million years) from mammalian species [27].

While the build number of the chicken genome is low compared to that of human and mouse, there are similar numbers of predicted proteins, but many fewer ESTs are available to aid in the structural annotation. We have used mass spectra from chicken serum to improve structural annotation of the chicken genome. We also address methods for database organization to obtain a significant number of peptide spectra matches when searching against databases derived from genetic sequences.

Results and Discussion

Our initial proteogenomic mapping experiments with several previously published chicken mass spectral datasets using a decoy database search strategy resulted in very few peptides that were unique to the genome when spectra were identified using searches against to the genomic database. Our decoy databases for the proteome and genome were derived using Markov chains based on the chicken proteome and the chicken genome respectively. Further investigation of the Δ CN and XCorr quality scores from the Sequest searches against the proteome and genome revealed that there were many high scoring peptide matches against the proteome, but very few high scoring matches against the genome as illustrated in Figure 5.1. Similar results were obtained for a chromosome relatively poor in serum protein genes (chr 6) and a chromosome relatively rich in serum protein genes (chr 3). Therefore we investigated different database construction approaches.

Since the size of the genome database is more than six times the size of the proteome database, we conducted an experiment to determine if the loss in peptide identifications against the genome was a function of database size. For this experiment, the proteome database was concatenated with increasing amounts of decoy random amino acid sequence to serve as the proteome database. Decoy databases of the same size were also generated and these databases were used to search for PSMs. Figure 5.2 shows the effects of database size on the number of peptides identified. The number of peptides identified against the proteome was used as a baseline and is indicated by zero on the x-axis (zero added decoy sequence) and one on the y-axis. Decoy amino acid

sequence was progressively added to the proteome in increments the same size as the proteome. Thus, a value of two on the x-axis means that the database was three times as large (proteome + 2x decoy) as the original proteome database. We noted a significant decrease in the peptide identifications as the number of spectra from decoy sequence increased. As Figure 5.2 indicates, there is substantial difference in the loss of peptide identifications among different datasets indicating differences in the quality of the mass spectra. Because these poor quality spectra are not robust to the addition of noise, they are not useful for proteogenomic mapping. This process of iteratively adding "noise" to the proteome and conducting searches provides a method for determining if spectra are of sufficiently high quality for proteogenomic mapping. As shown in Figure 5.2, the new serum proteomics dataset used for this study (collected using updated methodologies and equipment) is substantially more robust to the addition of noise than the older datasets.

The influence of database size on results obtained when searching against the genomic sequences translated in 6 reading frames also led us to use separate databases for genic and intragenic regions. In order to maintain a one-to-one relationship between the number of proteins in our protein database and the number of genes in our genic database, we constructed all three of our databases (proteomic, genic, and intragenic) based on the chicken proteins in the International Protein Index (IPI) database, which provides "minimally redundant yet maximally complete sets of proteins for featured species" [29]. Our genic database was generated by locating the gene sequences corresponding to proteins found in the IPI database, and extracting the DNA sequence including introns plus 5' and 3' UTR sequence from the sequenced genome. Since the

lengths of the 5' and 3' UTR sequences have not been determined experimentally for most chicken genes, we used 5' and 3' UTR lengths based on the mean lengths for *Homo sapiens* [30]). Peptides were identified using the Sequest spectral matching algorithm [31] by searching against our three databases and validated using a target-decoy database search strategy. Peptides with a p-value of less than 0.05 were considered valid identifications. Peptides identified by searching against the genic or intergenic databases but against the proteome database were analyzed with the Proteogenomic Mapping Tool [32] to generate the reverse translated (RT) sequences for each of the peptides by mapping each of these RT peptides back to locations within the chicken genome.

Our searches against the IPI *Gallus gallus* proteome database identified 268 proteins comprising the serum proteome. These proteins were identified by 8,797 peptides (960 unique peptides). We also identified 2,993 peptides when searching against our genic database. Of these 2,933 genic peptides, 2,742 were present in the results of the proteome search, resulting in 251 peptides (represented by 48 unique peptides) matching the genic database but not the proteome database. After examination of these peptide sequences, we identified 4 peptides in this dataset which were digests of peptides present in the results of the proteome DB search. This resulted in the identification of 44 unique peptides that map to the genic database but not to the proteome. Ten of these genic peptides were mapped into proteins identified by different proteomic peptides through our search of the proteome database. Searches against the database composed of intragenic regions yielded three unique peptides.

The ten unique peptides identified from our genic database search that also have corresponding evidence for expression from our proteome database search are shown mapped to the genome in Figures 5.3 - 5.6. Figure 5.3 shows an instance where a peptide maps to a chicken EST within the 5' UTR of an IPI protein similar to α-2-macroglobulin (IPI00599918) curated proteins on the same strand (aqua track). The peptide aligns with the NCBI gene model for this protein, but not the Ensembl gene model. Since the IPI proteins are heavily derived from the Ensembl gene models, this peptide was not present in the proteome database we performed our proteome search against, but we did observe several hits of different peptides against the proteome database for the IPI/Ensembl model, indicating the NCBI model is more accurate. Figure 5.4 shows six peptides mapping on the same strand to regions within the Immunoglobulin (IG) Lambda Chain Variable-1 Region in a region of high translation expression.

Some of these peptides have no exon or EST evidence, and we hypothesize that we are observing a sufficiently sensitive proteogenomic mapping to pick up the splicing changes in the exons of IG variable regions as part of the synthesis of immunoglobulin. Figure 5.5 shows a single peptide mapping to the 5' UTR of serum albumin (IPI00574195). Many peptides matching the proteome database confirm the expression of serum albumin. Given that this peptide is on the same strand as serum albumin, it is indicative of a potential new exon or gene.. Figure 5.6 shows a peptide mapping to the same strand as the 5' UTR of an uncharacterized protein (IPI00821912) which we identified as expressed through our searches of the proteome database. This peptide maps to a chicken

EST in this region, providing proteomic evidence confirming expression of the EST, and a possible correction of the annotation information for this uncharacterized protein.

The remaining 34 genic and 3 intragenic peptides can be divided into five groups based on where they map within the genome:

1. Peptide confirming protein expression
2. Peptide confirming exon from mRNA
3. Peptide indicating novel exon or gene
4. Peptide correcting exon boundary
5. Peptide in or near a repeat region

Figures 5.7 through 5.11 show examples of each of these groups. In Figure 5.7 two reverse translated (RT) peptides not seen in the set of peptides identified with the proteome database are shown. The genomic peptide IPI00579242 maps on the same strand as a gene model present in both the NCBI/Ensembl gene sets confirming protein expression, and providing evidence of a potential new exon within this protein. The second genomic peptide, IPI00580765 is shown mapped into a region between two genes in both the NCBI/Ensembl protein sets along the same strand indicative of a potential new small gene or exon in this region. Figure 5.8a shows an RT peptide confirming an NCBI gene where there is no Ensembl gene in the area and Figure 5.8b shows the same peptide using the UCSC genome browser confirming multiple mRNA evidence for an exon in the indicated region. Notably, the gene model from NCBI identified in a) is not indicated as having a higher confidence RefSeq gene model by b), meaning there is not a curated gene model for this gene in the NCBI database. Figure 5.9 a) shows 3 distinct RT peptides mapping to a novel chicken exon or gene while b) shows the same three peptides visualized with the UCSC genome browser in order to gain the conservation

track. These peptides are shown to map to an area of high conservation, providing evidence of a novel chicken exon or gene in this area. Figure 5.10 a) shows an RT peptide expressed within a known gene model along the same strand, near an exon and two repeat elements. Figure 5.10 b) shows this same peptide using the UCSC genome browser clearly mapping to a repeat region within the genome identified by RepeatMasker. Figure 5.11 a) shows an RT peptide that corrects an exon boundary in the structural annotation of the chicken genome of an Ensembl gene. It maps to the same strand to the edge of a known Ensembl gene, and b) shows that the peptide maps to a region of conservation at the 3' end of this gene model. There is no NCBI gene model for this protein (IPI00595493).

Conclusions

We have confirmed the expression 268 serum proteins from our *Gallus gallus* proteome database. The 47 remaining peptides that map uniquely to the genic and intragenic regions of the *Gallus gallus* genome were used to improve the structural annotation by confirming 2 exons predicted by mRNA, providing evidence of 17 novel exons or genes, showing evidence of the expression of 7 repeat regions, and providing 4 corrections to the boundaries of known exons. These peptides serve as a complementary method to traditional structural annotation methodologies, and for model organism genomes like chicken, that do not undergo the same level of refinement as human or mouse, provide annotation correction information that might not be otherwise readily available.

The use of proteogenomic mapping as a tool to aid in the structural annotation of genomes shows that even the most up-to-date *de novo* or homology based computational gene prediction misses or incorrectly annotates a number genes. Additionally, proteogenomic mapping provides proof that a given protein is actually translated and expressed in a tissue as opposed to the evidence of translation obtained using massively parallel next-generation sequencing technologies. As mass spectrometry techniques improve and the speed of matching spectra to peptides increases due to both improved algorithms and increased computational power, proteogenomic mapping should be increasingly utilized to provide and confirm structural annotations of eukaryotes.

Future work should focus on identifying areas of the genome where there is discrepancy between the NCBI and Ensembl gene model datasets and identifying peptides identified as expressed from those areas as part of an effort to improve computational tools for gene prediction. Additionally, efforts to include peptides with a lower probability of expression when identified using a decoy database strategy could be incorporated by anchoring regions with high probability expression peptides and then including lower probability peptides locally. Alternatively, our strategy of constructing the genomic database based on the protein database and not searching raw genomic sequence or performing extensive experimental manipulations allows regions were protein expression is observed to be easily identified and potentially used for construction of smaller databases supporting stepwise searches. When combined with an anchoring method, this could potentially provide higher coverage of peptides to the genome from a given experimental dataset.

Methods and Materials

Mass Spectrometry Datasets

The initial *Gallus gallus* serum MS dataset used is described in [33] and the *Gallus gallus* bursa MS dataset used for Figure 5.2 is described in [22]. For the updated serum dataset, a new serum sample from a serum pathogen free broiler chicken was taken.

Protein Isolation

Serum was collected from clotted whole chicken blood in inverted 3 cc/mL syringes. Serum was aliquoted, clarified (1000 rpm, 5 min, 4 °C) and protein yield was determined using the Pierce BCA Protein Assay Kit (Fisher Scientific, Pittsburgh, PA). One-dimensional (1-D) gel electrophoresis was performed on serial dilutions of the serum (1:2, 1:10, 1:25)(Criterion Gel System, Bio-Rad, Hercules, CA). Gels were stained with Coomassie Blue (Processor Plus, Amersham Biosciences, Piscataway, NJ) and documented (FluorChem SP, Alpha Innotech, San Leandro, CA).

Trypsin Digestion

In-gel and in-solution tryptic digestions on the serum were done in parallel for protein coverage comparison (43.4 μg protein were used). Prior to in-gel tryptic digestion, the lane representing the 1:10 dilution was selected and divided into 5 fractions: group 1 (darkest bands), group 2 (medium bands), group 3D (darkest light areas between bands), group 3M (medium light areas) and group 3L (lightest light areas).

Gel fractions were destained (50 mM NH_4HCO_3/50 % acetonitrile, 10 min), dehydrated (100 % acetonitrile, 15 min), reduced (10 mM DTT 30 min), alkylated (55 mM iodoacetamide, 20 min), dehydrated (100 % acetonitrile, 3 × 5 min), rehydrated (50 mM NH_4HCO_3, 10 min) and trypsin-digested (10 µg, O/N, Promega, Madison, WI). All steps were done at 35-37 °C. Peptides from the fractions were extracted (1% TFA, 2% acetonitrile, 2 × 30 min; the second time with 100 % acetonitrile). For the in-solution tryptic digestions, aliquots (10 µL) of the 1:10 dilution were reduced (5 mM DTT, 5 min, 65 °C), alkylated (10 mM iodoacetamide, 30 min, 37 °C) and trypsin-digested (1 µg, O/N, 37 °C, Promega, Madison, WI). All in-gel and in-solution tryptic digests were vacuum centrifuged until dried completely (Savant SPD2010, Thermo Electron, Milford, MA) and resuspended in 0.1 % formic acid. Digests from groups 3D, 3M and 3L were pooled, vacuum centrifuged and resuspended in 0.1 % formic acid.

Sample Cleanup

After digestion, samples are adjusted to 2% Acetonitrile and each is desalted using a peptide macrotrap (Michrom TR1/25108/52) according the manufacturer's instructions. Following desalting, samples are cleaned using a strong cation exchange (SCX) trap (Michrom TR1/25108/53) according to the manufacturer's instructions to remove detergents or other polymers which can interfere with MS/MS analysis. All samples were then dried and resuspended in 20 µl of 5% Acetonitrile, 0.1% Formic Acid and transferred to a low retention autosampler vial for deconvolution via reverse phase, high pressure liquid chromatography.

Nanospray LC/MS

Each sample was loaded on a BioBasic C18 reversed phase column (Thermo 72105-100266) and flushed for 20 min with 5% acetonitrile (ACN), 0.1% formic acid to remove salts. Peptide separation was achieved using a Thermo Surveyor MS pump with a 655 min nano-HPLC method consisting of a gradient from 5% ACN to 50% ACN in 620 min, followed by a 20 minute wash with 95% ACN and equilibration with 5% ACN for 15 minutes (all solvents contain 0.1% formic acid as a proton source). Ionization of peptides was achieved via nanospray ionization using a Thermo Finnigan nanospray source type I operated at 1.85kV with 8 micrometer internal diameter silica tips (New Objective FS360-75-8-N-20-C12). High voltage was applied using a t-connector with a gold electrode in contact with the HPLC solvent. A Thermo LCQ DECA XP Plus ion trap mass spectrometer was used to collect data over the 655 minute duration of each HPLC run. Precursor mass scans were performed using repetitive MS scans, each immediately followed by three MS/MS scans of the three most intense MS peaks. Dynamic exclusion was enabled with a duration of two minutes and a repeat count of two.

Protein Identification

Database searches were performed using the SEQUEST algorithm [34] in Bioworks 3.3 (Thermo Finnigan). The proteome database for peptide spectral matching was the *Gallus gallus* IPI protein database (version 3.56) [29]. Search results were filtered using a decoy based, distance-based outlier detection method in which a

probability of being a false positive match is assigned to each peptide [35]. The decoy database was constructed using a 0^{th} order hidden Markov model based on the amino acid distribution of the proteome. Proteins with a probability of differential expression of 0.05 or less were selected for further modeling. Two different databases for peptide spectral matching against the genome were constructed based on the *Gallus gallus* IPI protein database (version 3.56). The first database contains the genetic sequences of all the proteins contained within the IPI protein set including introns and 5' and 3' regulatory regions [30], and the second database contains all the remaining intragenic regions not contained in the first database. Decoy databases were constructed using a 5^{th} order hidden Markov model based on the nucleotide distribution in each of these two databases.

Proteogenomic Mapping

Custom Perl scripts were used to identify peptides with unique peptide spectral matches to the genomic databases and these peptides were mapped onto the chicken genome using the Proteogenomic Mapping Tool, a Java based tool which implements the Aho-Corasick string mapping algorithm for proteogenomic mapping in both prokaryotes and eukaryotes [32]. The resulting output was then visualized using Gbrowse [36, 37] and the UCSC Genome Browser [38], and GFF3 files were also generated for use with other genome browsers.

LITERATURE CITED

1. Mathe C, Sagot MF, Schiex T, Rouze P: **Current methods of gene prediction, their strengths and weaknesses**. *Nucleic Acids Res* 2002, **30**(19):4103-4117.

2. Allen JE, Pertea M, Salzberg SL: **Computational gene prediction using multiple sources of evidence**. *Genome Res* 2004, **14**(1):142-148.

3. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L: **Identification and correction of abnormal, incomplete and mispredicted proteins in public databases**. *BMC Bioinformatics* 2008, **9**(353):353.

4. Castellana N, Bafna V: **Proteogenomics to discover the full coding content of genomes: a computational perspective**. *J Proteomics* 2010, **73**(11):2124-2135.

5. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Matching peptide mass spectra to EST and genomic DNA databases**. *Trends Biotechnol* 2001, **19**(10 Suppl):S17-22.

6. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation**. *Proteomics* 2004, **4**(1):59-77.

7. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD: **Proteogenomics: needs and roles to be filled by proteomics in genome annotation**. *Brief Funct Genomic Proteomic* 2008, **7**(1):50-62.

8. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N *et al*: **The complete genome and proteome of Mycoplasma mobile**. *Genome Res* 2004, **14**(8):1447-1461.

9. Savidor A, Donahoo RS, Hurtado-Gonzales O, Verberkmoes NC, Shah MB, Lamour KH, McDonald WH: **Expressed peptide tags: an additional layer of data for genome annotation**. *J Proteome Res* 2006, **5**(11):3048-3058.

10. Wilkins MJ, Verberkmoes NC, Williams KH, Callister SJ, Mouser PJ, Elifantz H, N'Guessan A L, Thomas BC, Nicora CD, Shah MB *et al*: **Proteogenomic monitoring of Geobacter physiology during stimulated uranium bioremediation**. *Appl Environ Microbiol* 2009, **75**(20):6591-6599.

11. Kunec D, Nanduri B, Burgess SC: **Experimental annotation of channel catfish virus by probabilistic proteogenomic mapping**. *Proteomics* 2009, **9**(10):2634-2647.

12. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C, Lecompte O: **Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol**. *Genome Res* 2009, **19**(1):128-135.

13. Nanduri B, Wang N, Lawrence ML, Bridges SM, Burgess SC: **Gene model detection using mass spectrometry**. *Methods Mol Biol* 2010, **604**:137-144.

14. Payne SH, Huang ST, Pieper R: **A proteogenomic update to Yersinia: enhancing genome annotation**. *BMC Genomics* 2010, **11**(460):460.

15. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S *et al*: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry**. *Genome Biol* 2005, **6**(1):R9.

16. Matis M, Zakelj-Mavric M, Peter-Katalinic J: **Mass spectrometry and database search in the analysis of proteins from the fungus Pleurotus ostreatus**. *Proteomics* 2005, **5**(1):67-75.

17. Smith JC, Northey JG, Garg J, Pearlman RE, Siu KW: **Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of Tetrahymena thermophila**. *J Proteome Res* 2005, **4**(3):909-919.

18. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ: **Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics**. *Genome Biol* 2006, **7**(4):R35.

19. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP: **Discovery and revision of Arabidopsis genes by proteogenomics**. *Proc Natl Acad Sci U S A* 2008, **105**(52):21034-21038.

20. Colinge J, Cusin I, Reffas S, Mahe E, Niknejad A, Rey PA, Mattou H, Moniatte M, Bougueleret L: **Experiments in searching small proteins in unannotated large eukaryotic genomes**. *J Proteome Res* 2005, **4**(1):167-174.

21. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry**. *Genome Res* 2007, **17**(2):231-239.

22. McCarthy FM, Cooksey AM, Wang N, Bridges SM, Pharr GT, Burgess SC: **Modeling a whole organ using proteomics: the avian bursa of Fabricius**. *Proteomics* 2006, **6**(9):2759-2771.

23. Kalume DE, Peri S, Reddy R, Zhong J, Okulate M, Kumar N, Pandey A: **Genome annotation of Anopheles gambiae using mass spectrometry-derived data**. *BMC Genomics* 2005, **6**(128):128.

24. Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, Stephenson JL, Jr.: **Whole genome searching with shotgun proteomic data: applications for genome annotation**. *J Proteome Res* 2008, **7**(1):80-88.

25. Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ: **Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations**. *Genome Res* 2008, **18**(10):1660-1669.

26. Lucitt MB, Price TS, Pizarro A, Wu W, Yocum AK, Seiler C, Pack MA, Blair IA, Fitzgerald GA, Grosser T: **Analysis of the zebrafish proteome during embryonic development**. *Mol Cell Proteomics* 2008, **7**(5):981-994.

27. **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 2004, **432**(7018):695-716.

28. Smith J, Bruley CK, Paton IR, Dunn I, Jones CT, Windsor D, Morrice DR, Law AS, Masabanda J, Sazanov A *et al*: **Differences in gene density on chicken macrochromosomes and microchromosomes**. *Anim Genet* 2000, **31**(2):96-103.

29. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments**. *Proteomics* 2004, **4**(7):1985-1988.

30. Hong X, Scofield DG, Lynch M: **Intron size, abundance, and distribution within untranslated regions of genes**. *Mol Biol Evol* 2006, **23**(12):2392-2404.

31. Eng JK, McCormack AL, Yates JR, 3rd: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *Analytical chemistry* 1994, **5**:976-989.

32. Sanders WS WN, Bridges SM, Malone BM, Dandass YS, McCarthy FM, Nanduri B, Lawrence M, Burgess SC: **The Proteogenomic Mapping Tool**. *BMC Bioinformatics* 2011, **(accepted)**.

33. Corzo A, M. T. Kidd, G. T. Pharr, and S. C. Burgess.: **Initial mapping for the chicken blood plasma proteome.** *International Journal of Poultry Science* 2004, **3**:157-162.

34. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database**. *Anal Chem* 1995, **67**(8):1426-1436.

35. Wang N, Yuan C, Wu D, Nanduri B, SM B: **PepOut: a Distance-based Outlier Detection for Improving MS/MS Peptide Identification Confidence.** *International Journal of data mining and bioinformatics.* 2010, **(accepted)**.

36. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The generic genome browser: a building block for a model organism system database Gallus GBrowse: a unified genomic database for the chicken**. *Genome Res* 2002, **12**(10):1599-1610.

37. Schmidt CJ, Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J, McGrath S, Mardis E, Stein LD: **Gallus GBrowse: a unified genomic database for the chicken**. *Nucleic Acids Res* 2008, **36**(Database issue):D719-723.

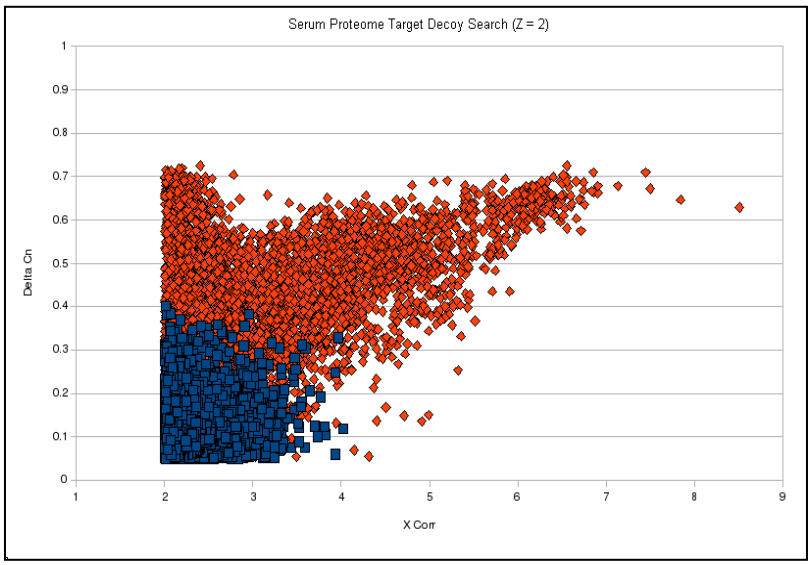
38. Karolchik D, Hinrichs AS, Kent WJ: **The UCSC Genome Browser**. *Curr Protoc Bioinformatics* 2009, **Chapter 1**(4):Unit1 4.

TABLE 5.1

COMPARISION OF GENOME SIZES FOR SELECTED PROTEOGENOMIC MAPPING PROJECTS

Organism	Genome Size (Mbp)	Reference
Channel catfish herpesvirus (CCV)	0.13	[11]
Mycoplasma mobile	0.78	[8]
Mycoplasma pneumonia	0.8	[5]
Haemophilus influenzae	1.8	[5]
Porphyromonas gingivalis	2.2	[5]
Haemophilus somnus	2.3	[13]
Pasteurella multocida	2.5	[13]
Mannheimia haemolytica	2.8	[13]
Geobacter lovleyi	3.9	[10]
Geobacter bemidjiensis	4.6	[10]
Yersinia pestis	4.7	[14]
Rhodopseudomonas palustris	5.5	[9]
Arabodopsis thaliana	~125	[19]
Caenorhabditis elegans	~100	[25]
Tetrahymena thermophila	~102	[17]
Anopheles gambiae	~278	[23]
Gallus gallus	~1500	[22]
Danio rerio	~3112	[26]
Homo sapiens	~3400	[15, 18, 20, 21, 24]

a)



b)

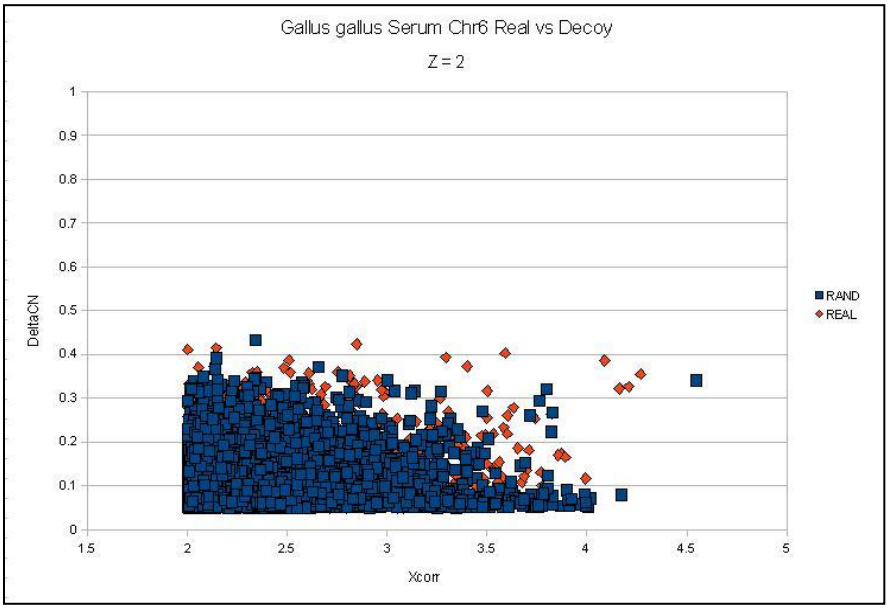


Figure 5.1

Initial Comparision of Peptide Spectra Matches Against the Proteome and *Gallus gallus* Chromosome 6.

a) *Gallus gallus* Proteome Target-Decoy Analysis
b) *Gallus gallus* Chromosome 6 Target-Decoy Analysis

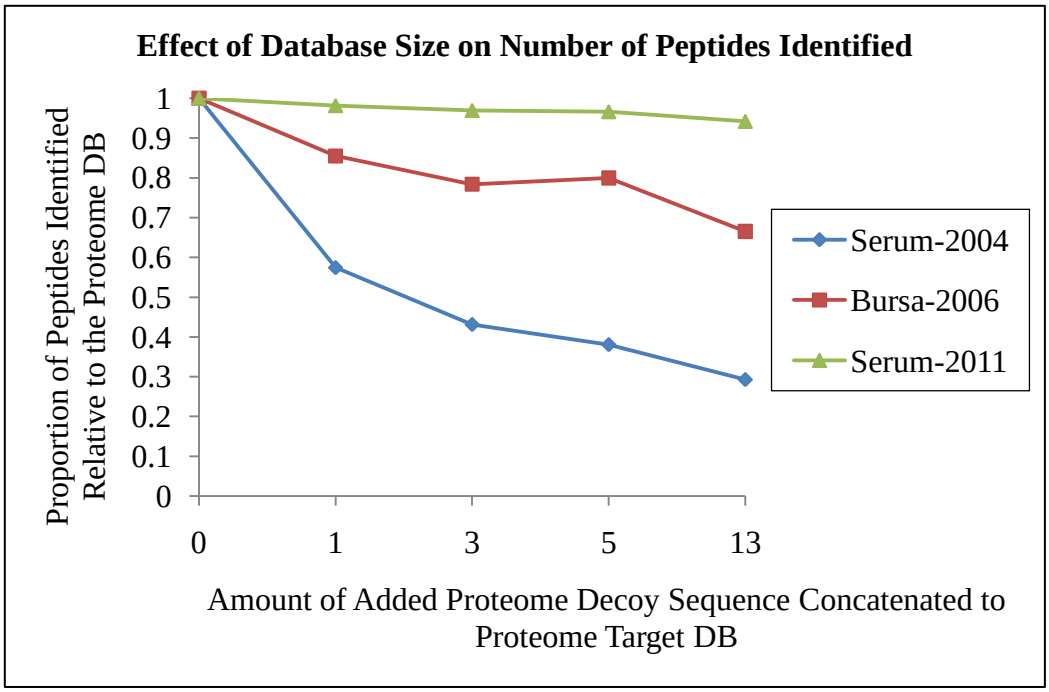


Figure 5.2

Loss In Shared Peptide Identifications
Between Proteome and Databases of Increasing Size

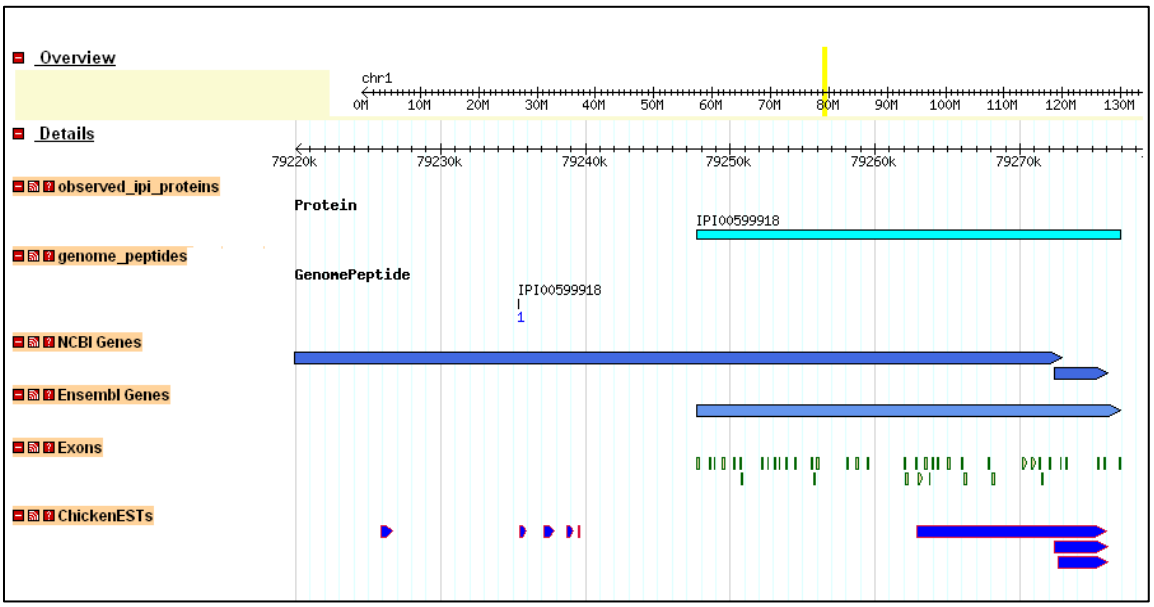


Figure 5.3

IPI00599918 – Similar to Alpha-2-Macroglobulin

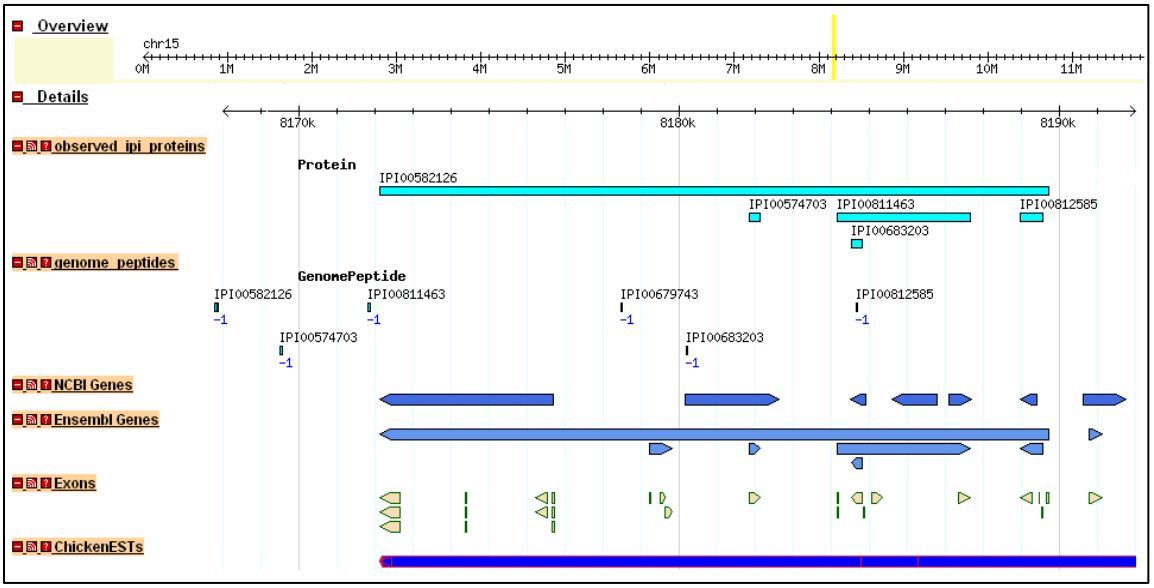


Figure 5.4

IPI00582126 – IG Lambda Chain V-1 Region

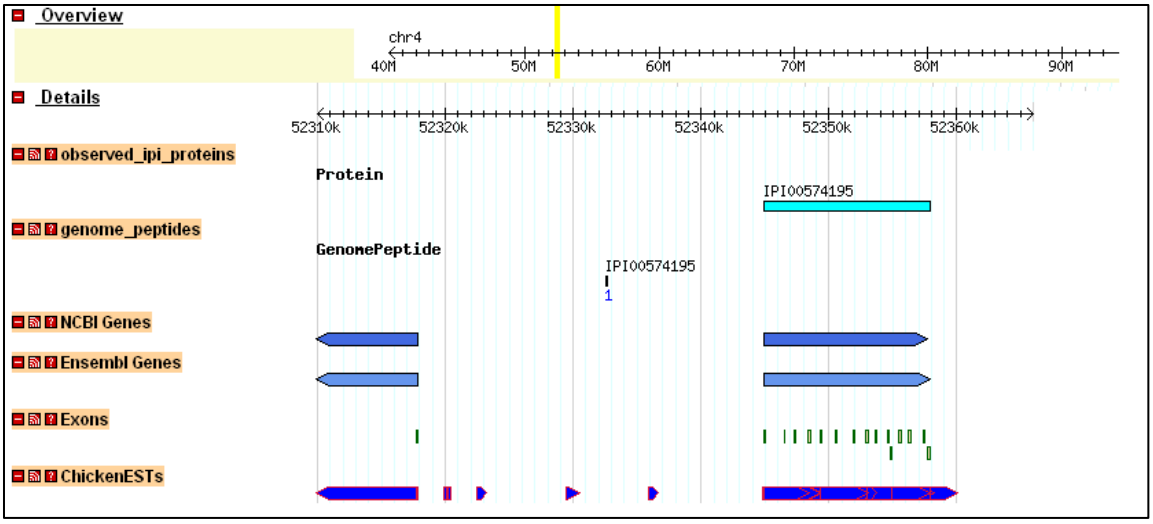


Figure 5.5

IPI00574195 – Serum Albumin

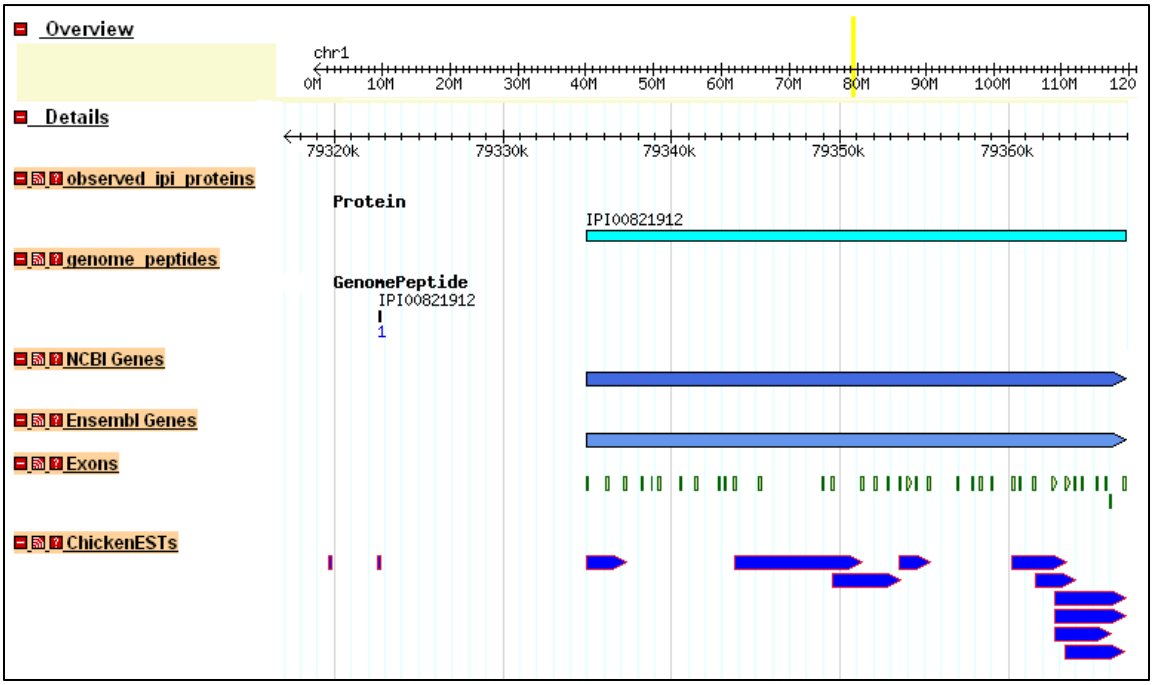


Figure 5.6

IPI00821912 – Uncharacterized Protein

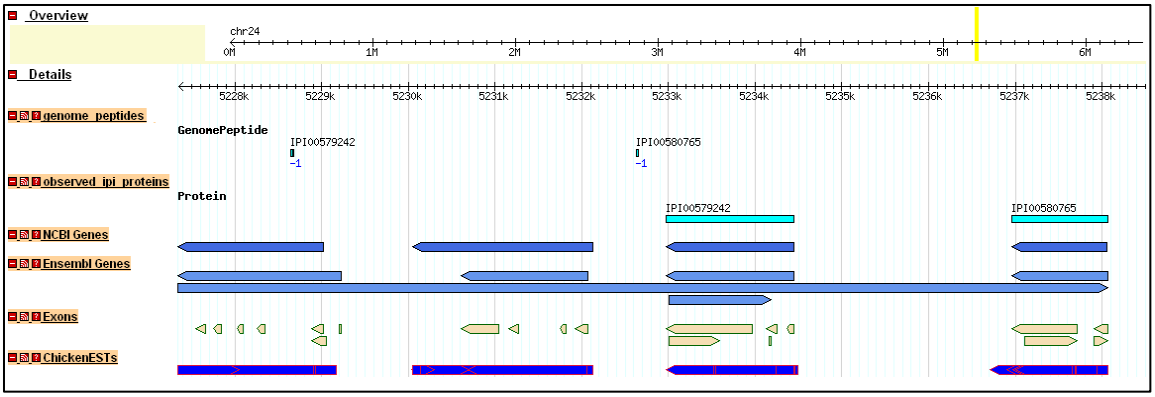
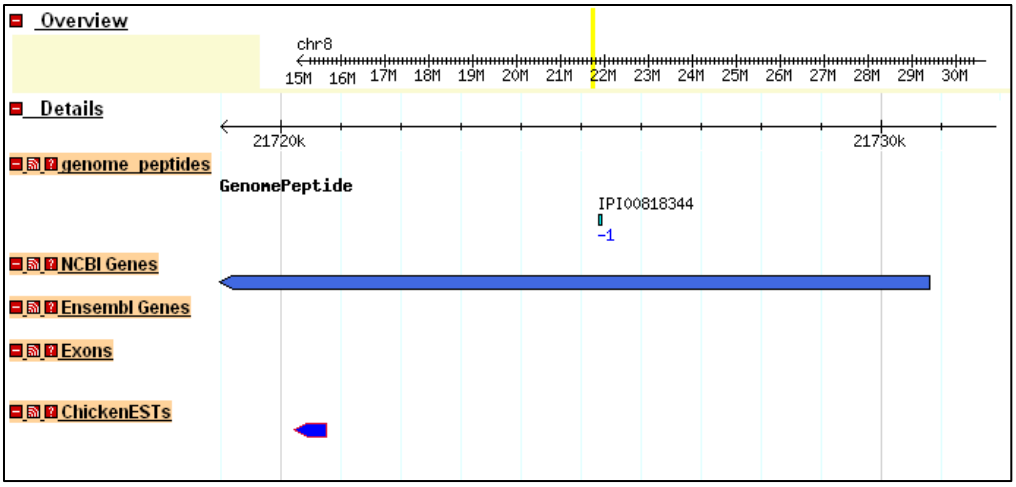


Figure 5.7

A Peptide Confirming Protein Expression and Possible Novel Exon and a Peptide Representing Novel Exon or Gene

a)



b)

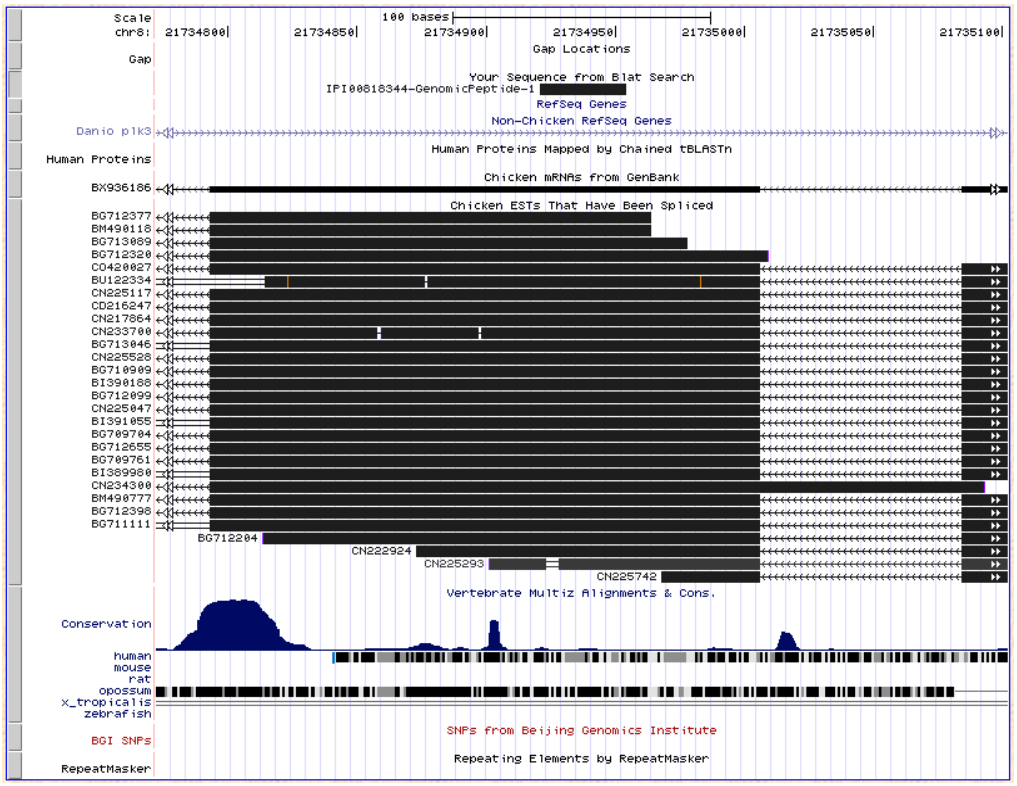
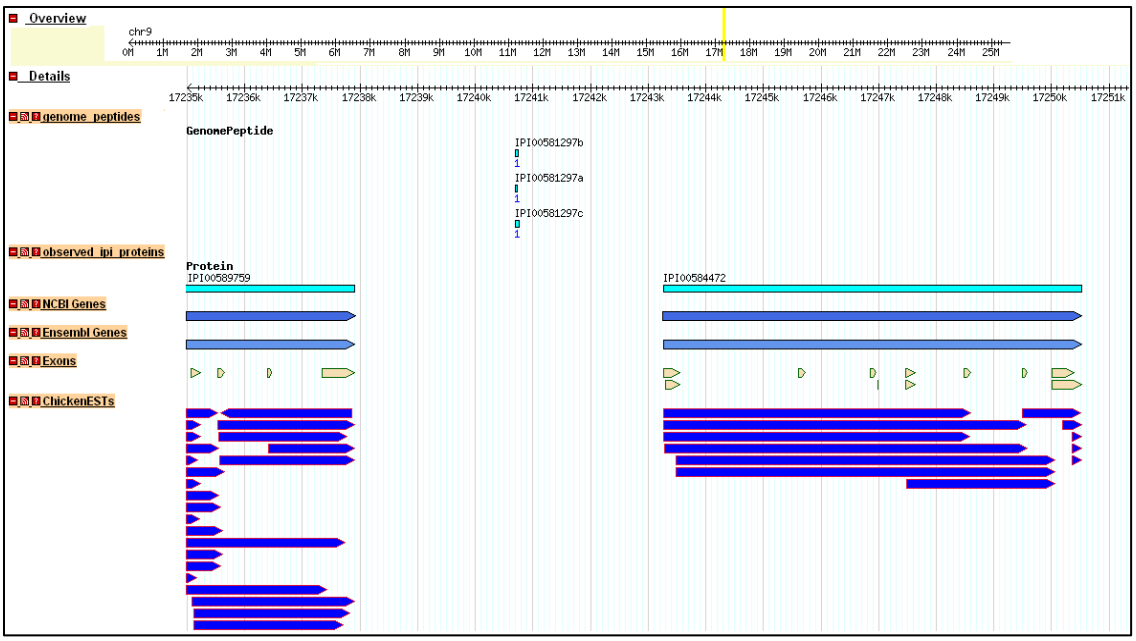


Figure 5.8

Peptide Confirming Exon From mRNA

a)



b)

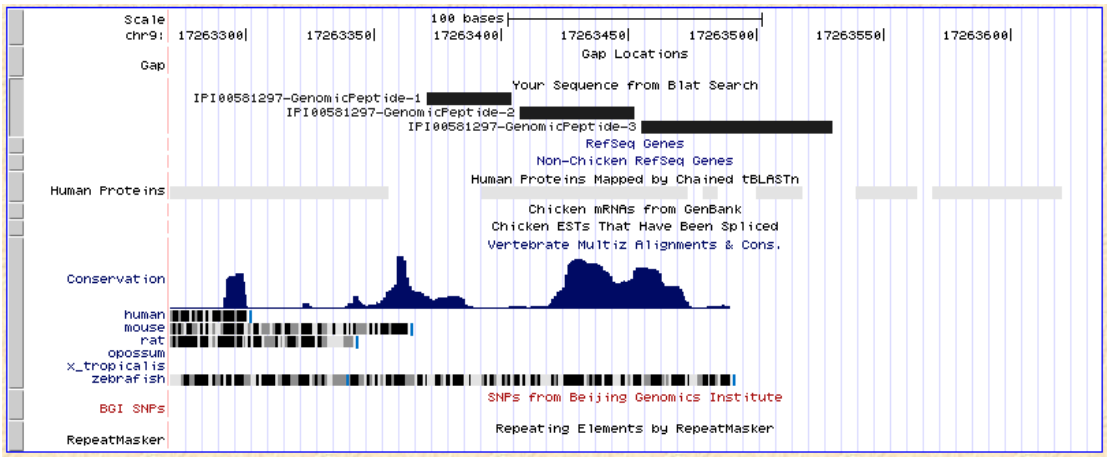
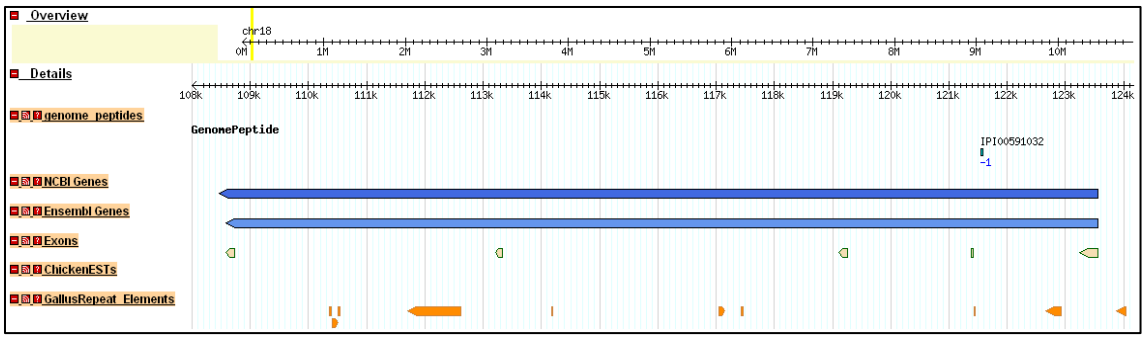


Figure 5.9

Peptide Indicating Novel Exon or Gene

a)



b)

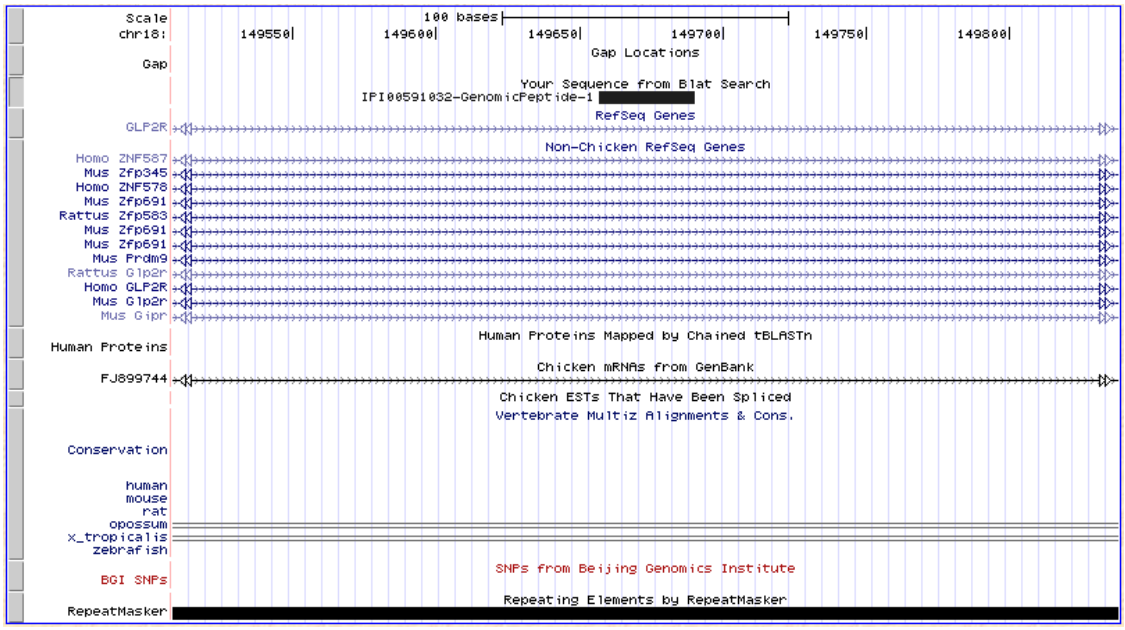
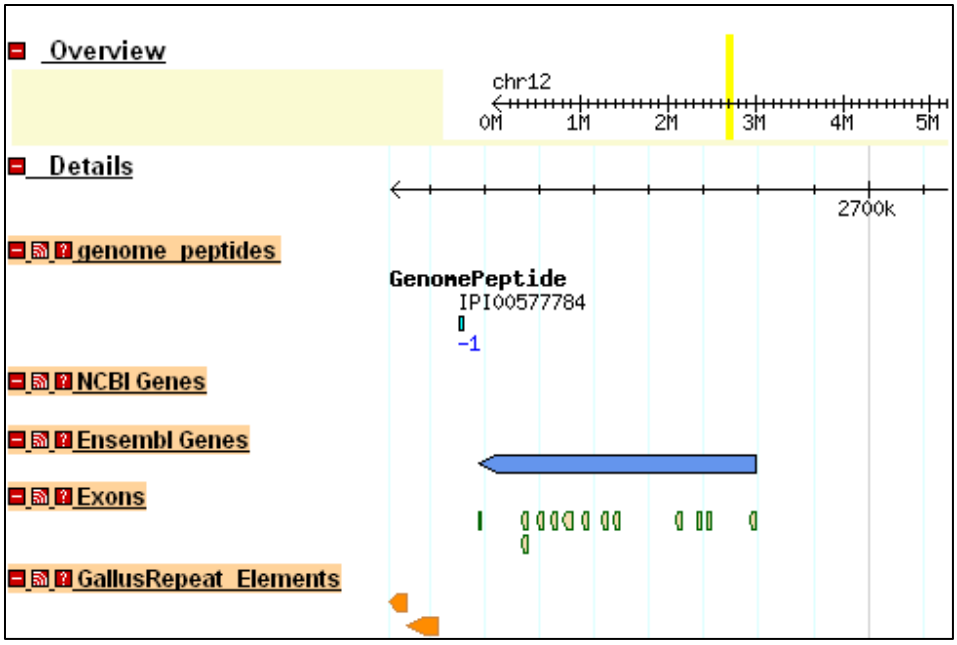


Figure 5.10

Peptide In or Near a Repeat Region

a)



b)

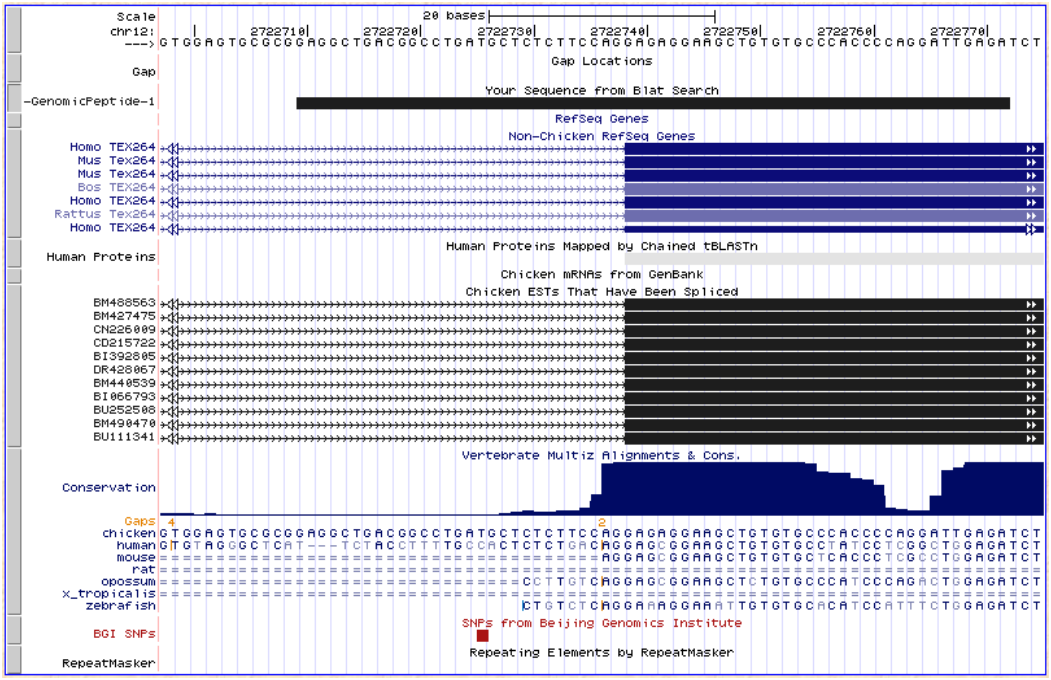


Figure 5.11

Peptide Correcting Exon Boundary

CHAPTER VI

CONCLUSIONS

The increasing volumes of genomic, transcriptomic and proteomic data available has resulted in the need for rapid analytical techniques that derive information from that data. This dissertation addresses the application of machine learning algorithms to proteomics problems. This chapter summarizes each project and evaluates the research as a whole. We have shown that it is possible to predict experimentally observable properties of peptides using machine learning classifiers and that the application of mass spectrometry derived proteomics data can help improve the structural annotation of genomes.

Prediction of Peptides Observable by Mass Spectrometry

Chapter II describes a procedure for constructing an artificial neural network classifier to predict which tryptic peptides in a protein are likely to be detectable by mass spectrometry for a specific set of experimental and instrumental conditions. We demonstrate that it is possible to construct a classifier with accuracy comparable to those previously reported based on the accumulation of large training sets from multiple experiments. We also show that a classifier constructed based on one dataset does not perform at an acceptable level when predicting observability for another dataset and thus

it is necessary to construct classifiers that are specific for one set of experimental conditions. The resulting classifier provides researchers with a tool that can provide information about peptide coverage of proteins in terms of which proteins are likely to be detectable. It can also be used as one line of evidence in a systems analysis to evaluate alternative hypotheses concerning proteins that were not observed but that were expected. If the "missing" protein generates many predicted detectable peptides but none were observed, then this provides additional probabilistic evidence of absence of the protein—a very difficult hypothesis to demonstrate conclusively. The classifier allows researchers to distinguish between proteins that are not likely to be detected with the methodology versus proteins that were not expressed in the biological system and to thus improve biological modeling.

Prediction of Cell Penetrating Peptides

We have identified sets of known cell penetrating peptides and non-penetrating cell penetrating analogs from the literature and use these to construct a number of different datasets to address the problem of imbalance between the number of positive and negative examples. Utilizing these datasets, we show that it is possible to obtain a higher than previously reported accuracy for the prediction of cell penetrating peptides using support vector machines as opposed to previous methods utilizing a method based on determining if the average score of a peptide falls within a range of features determined through the use of principle component analysis. We then generate a number of peptides based on the amino acid distribution of the chicken proteome and classify

these peptides as either cell penetrating or non-cell penetrating based on the predictions of our classifiers. These peptides along with positive and negative experimental controls were synthesized and analyzed for cell penetration using two avian cell lines. Our classifiers accurately predict cell-pentrating potential, and we have identified a lack of negative examples of cell penetrating peptides in the literature. Additionally, we have noted that the cell type being used for the evaluation of cell penetrating potential should be included as a predictive feature in future studies as peptides previously predicted to be penetrating or non-penetrating in previous studies using a specific cell line might not be accurate for alternative cell lines.

Proteogenomic Mapping of Chicken Serum

We have confirmed the expression 268 serum proteins from our *Gallus gallus* proteome database. The 47 remaining peptides that map uniquely to the genic and intragenic regions of the *Gallus gallus* genome were used to improve the structural annotation by confirming 2 exons predicted by mRNA, providing evidence of 17 novel exons or genes, showing evidence of the expression of 7 repeat regions, and providing 4 corrections to the boundaries of known exons. These peptides serve as a complimentary method to traditional structural annotation methodologies, and for model organism genomes like chicken, that do not undergo the same level of refinement as human or mouse, provide annotation correction information that might not be otherwise readily available.

The use of proteogenomic mapping as a tool to aid in the structural annotation of genomes shows that even the most up-to-date *de novo* or homology based computational gene prediction misses or incorrectly annotates a number genes. Additionally, proteogenomic mapping serves provides proof that a given protein is actually translated and expressed in a tissue as opposed to the evidence of translation obtained using 2nd generation sequencing technologies. As mass spectrometry techniques improve and the speed of matching spectra to peptides increases due to both improved algorithms and increased computational power, proteogenomic mapping should be increasingly utilized to provide and confirm structural annotations of eukaryotes.

Future work should focus on identifying areas of the genome where there is discrepancy between the NCBI and Ensembl gene model datasets and identify any peptides identified as expressed from those areas as part of an effort to improve computational tools for gene prediction. Additionally, efforts to include peptides with a lower probability of expression when identified using a decoy database strategy could be incorporated by anchoring regions with high probability expression peptides and then including lower probability peptides locally. Alternatively, our strategy of constructing the genomic database based on the protein database and not searching raw genomic sequence or performing extensive experimental manipulations allows regions where protein expression is observed to be easily identified and potentially used for construction of smaller databases to search against in a stepwise manner. When combined with an anchoring method, this could potentially provide higher coverage of peptides to the genome from a given experimental dataset.

Summary

In Chapter II, we have shown that it is possible to predict the peptides observable by mass spectrometry using neural networks for a single dataset in contrast to other prediction methods utilizing large datasets compiled from a number of different experimental techniques. Since the research presented in that chapter was conducted and published, methods for statistically validating peptides from mass spectrometry derived datasets have changed to provide a more statistical basis for determining which peptides are valid hits against a target proteome database. Given these changes, future work in predicting peptides observable by mass spectrometry could explore different strategies for training dataset construction methodologies. Additionally, to obtain a better understanding of which properties contribute to the prediction of flyability could be obtained by analyzing the same proteomics mixture using a variety of digestion enzymes in addition to trypsin while holding the LC-MS conditions constant. The entire database of theoretical spectra could be analyzed to determine which peptides are proteotryptic, reducing the size of the database to be searched against, and increasing the effectiveness proteogenomic mapping for organisms with large genomes.

The research presented in Chapter III shows that machine learning algorithms using individual biochemical properties for features instead of composite features derived from principle component analysis can accurately predict whether or not a given peptide is capable of cell penetration. The small amount of data available for constructing training and testing datasets were discussed, and demonstrates the critical need for a curated database of peptides shown to penetrate or not penetrate a given cell type.

Additionally, future research should focus on compiling sets of non-penetrating analogs of cell penetrating peptides to increase the number of difficult to predict negative examples and to aid in the improvement of the prediction accuracy of classifiers. Additionally, once individual cell types are included in the features used for prediction of cell penetration, various features such the lipid composition of the cell membranes of various cell lines could be included to aid in prediction and help provide better understanding of the mechanism of cell penetration for that cell type.

Computational chemistry has provided software packages to analyze the quantitative structure and activity relationships (QSAR) of molecules, and these QSAR features were shown by Dobachev et al. [1] to aid in the prediction of cell penetrating peptides, and could be combined with the basic biochemical properties we utilize for features to improve the prediction accuracy of our classifiers for the prediction of flyable peptides and cell penetration.

Chapter IV presents a tool for taking peptides observed via mass spectrometry and mapping them back to the genome in a fast and accurate manner in order to help improve the structural annotation of genomes. In Chapter V we utilize this method along with various methods for assessing dataset quality and database construction to take peptides observed from the serum of the domestic chicken and use them for proteogenomic mapping to improve the structural annotation of the chicken genome. We show that proteogenomic mapping is sufficiently sensitive to identify variations in splicing events used to produce various immunoglobulin isoforms, and identify corrections to intron/exon boundaries in predicted genes, and identify novel genes and exons not

identified by traditional structural genome annotation methods. As more eukaryotic proteogenomic mapping projects make progress, these novel genes, exons, and boundary corrections should be incorporated in general gene models for improved gene predictions and used for the correction of structural annotations in genomes.

LITERATURE CITED

1. Dobchev DA, Mager I, Tulp I, Karelson G, Tamm T, et al. Prediction of Cell-Penetrating Peptides Using Artificial Neural Networks. Curr Comput Aided Drug Des 2010: 6.