Research statement

William Stafford Noble

The trend in biology toward the development and application of high-throughput, genome- and proteome-wide assays necessitates an increased reliance upon computational techniques to organize and understand the results of biological experiments. Without appropriate computational tools, biologists cannot hope to fully understand, for example, a complete genome sequence or a collection of hundreds of thousands of mass spectra. My research focuses on the development and application of methods for interpreting complex biological data sets. These methods may be used, for example, to uncover distant structural and functional relationships among protein sequences, to identify transcription factor binding site motifs, to classify cancerous tissues on the basis of microarray mRNA expression profiles, to predict properties of local chromatin structure from a given DNA sequence, and to accurately map tandem mass spectra to their corresponding peptides.

The goals of my research program are to develop and apply powerful new computational methods to gain insights into the molecular machinery of the cell. In selecting research areas to focus on, I am drawn to research problems in which I can solve fundamental problems in biology and human disease while also pushing the state of the art in machine learning.

Pattern recognition in diverse and heterogeneous genomic and proteomic data sets

Genome sciences is, in many ways, a data-driven enterprise because available technologies define the types of questions that we can ask. Each assay—DNA sequencing, the yeast two-hybrid screen, tandem mass spectrometry —provides one view of the molecular activity within the cell. An ongoing theme in my research is the integration of heterogeneous data sets, with the aim of providing a unified interpretation of the underlying phenomenon. We focus, in particular, on inferring gene function and on predicting protein-protein interactions. For example, to determine whether a given target pair of proteins interact, we take into account direct experimental evidence in the form of a yeast two-hybrid assay or tandem affinity purification followed by mass spectrometry. In addition, we consider as evidence the sequence similarity between the target pair of proteins and one or more pairs of proteins that are known to interact with one another, the similarity of the target proteins' mRNA expression profiles or ChIP-chip expression profiles, and evidence of cellular colocalization. We have developed a statistical inference framework that considers all of these sources of evidence, taking into account dependencies among them and weighting each type of evidence according to its relevance and its trustworthiness.

Much of my research program relies on two complementary classes of methods. The first class of methods, developed recently in machine learning, are known as *kernel methods* [78]. An algorithm is a kernel method if it relies on a particular type of function (the *kernel function*) to define similarities between pairs of objects. For these algorithms, a data set of N objects can be sufficiently represented using an N-by-N matrix of kernel values. The kernel matrix thereby provides a mechanism for representing diverse data types using a common formalism.

In collaboration with a variety of research groups, we have demonstrated the broad applicability of kernel methods to problems in genomics and proteomics, focusing on a particular

1	metabolism
2	energy
3	cell cycle & DNA processing
4	transcription
5	protein synthesis
6	protein fate
7	cellular transp. & transp. mech.
8	cell rescue, defense & virulence
9	interaction w/ cell. envt.
10	cell fate
11	control of cell. organization
12	transport facilitation
13	others

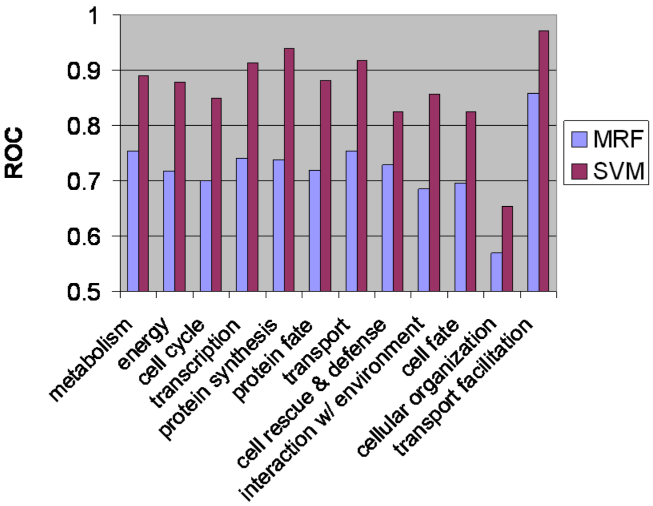

Figure 1: **Predicting yeast gene function from heterogeneous data.** The height of each bar is proportional to the cross-validated receiver operating characteristic score for prediction of a given class of yeast genes. The figure compares the performance of a previously published Markov random field method (in red) [20] and two variants of our SVM-based method (yellow and green). In every case, the SVM significantly outperforms the MRF [49].

kernel method known as the support vector machine (SVM) [9]. The SVM is a kernel-based classification algorithm that boasts strong theoretical underpinnings [85] as well as state-of-the-art performance in a variety of bioinformatics applications [58]. We have shown that

- SVMs can successfully classify yeast genes into functional categories on the basis of microarray expression profiles [11] or motif patterns within promoter sequences [65, 86].
- SVMs can discriminate with high accuracy among subtypes of soft tissue sarcoma on the basis of microarray expression profiles [80, 79]. Our SVM classifier provided strong evidence for several previously described histological subtypes, and suggested that a subset of one controversial subtype exhibits a consistent genomic signature.
- A series of SVM-based methods can recognize protein folds and remote homologs [52, 50, 51, 88, 36, 55]. Our early work in this area set the baseline against which much subsequent work was compared, including many SVM-based classifiers that derive from our work [5, 46, 12, 23, 62, 63, 72, 77, 47].
- SVMs can be applied to a variety of applications within the field of tandem mass spectrometry, including re-ranking peptide-spectrum matches produced by a database search algorithm [1, 37] and assigning charge states to spectra [45].
- SVMs can draw inferences from heterogeneous genomic and proteomic data sets. We first demonstrated how to infer gene function from a combination of microarray expression profiles and phylogenetic profiles [66], and we subsequently described a statistical

framework for learning relative weights for each data set with respect to a given inference task [49, 48] (see Figure 1). We have also used this framework to predict protein-protein interactions [6] and protein co-complex relationships [71] from heterogeneous data sets.

The SVM is now one of the most popular methods for the analysis of biological data sets: Pubmed includes 387 papers published within the last 12 months whose abstracts contain the phrase "support vector machine," and 1488 such papers in the last five years. *Nature Biotechnology* invited me to write a primer on SVMs [59]. My research bears considerable responsibility for the SVM's popularity, because I have repeatedly demonstrated the power and flexibility of this algorithm in new bioinformatics domains.

The second class of methods that we use regularly is the Bayesian network. A Bayesian network is a formal graphical representation of a joint probability distribution over a collection of random variables. We have made particular use of dynamic Bayesian networks (DBNs) for modeling time series data, and a specific type of DBN known as the hidden Markov model (HMM). Starting with my PhD research, I have used HMMs for modeling motifs in DNA and protein sequences [31, 3]. More recently, we have used DBNs to model peptide fragmentation in a mass spectrometer [44], transmembrane protein topology [75], DNA-binding footprints in DNaseI sensitivity data [14] and nucleosome positioning signals in genomic DNA [73]. Compared with discriminative modeling methods such as the SVM, a Bayesian network offers several important advantages, including allowing a principled method for handling missing data, providing a complementary means of encoding prior knowledge, and providing a model that gives explanations for its predictions.

My lab will continue to apply these two complementary modeling approaches, both separately and jointly, to various applications. In particular, we are interested in coupling these core learning strategies with new ideas from the field of machine learning. These include, for example, using *semisupervised learning* [13] to leverage unlabeled data, *metric space embedding* [2] and *deep learning* [7, 17] to automatically ascertain structure in a rich set of features, and *multitask learning* [18] to exploit hidden dependencies among related learning tasks. For example, we have recently developed a deep neural network architecture that is trained in a multitask fashion to predict multiple local properties, including secondary structure, solvent accesibility, transmembrane topology, signal peptides and DNA-binding residues. The method provides state-of-the-art performance on all of these tasks, thus providing a unified framework for characterizing local protein properties. We plan to adapt similar strategies for characterizing chromatin structure and for analyzing mass spectrometry data.

The relationships among primary DNA sequence, chromatin and genome structure

DNA in the nucleus of the cell is bound in a complex and dynamic molecular structure known as chromatin. Chromatin structure, from the local scale up to the global 3D structure of chromosomes in the nucleus, has profound influences on gene regulation, DNA replication and repair, mutation and breakpoints. Over the past several years, my research group has investigated the relationships among the primary DNA sequence, nucleosomes, *cis*-regulatory factors, higher-order chromatin structure and the 3D structure of the genome. Initially, we

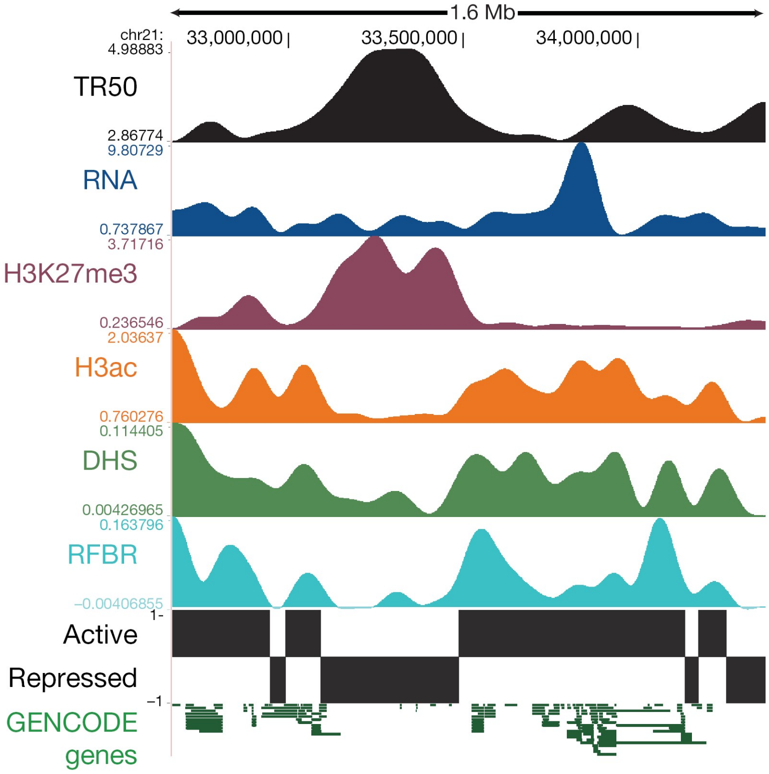

Figure 2: **Concordance of multiple data types for an illustrative ENCODE region (ENM005).** The tracks labeled "Active" and "Repressed" are derived from a simultaneous HMM segmentation of eight data types: replication time (TR50), bulk RNA transcription (RNA), histone modifications H3K27me3 and H3ac, DHS density and regulatory factor binding region density (RFBR).

focused on local disruptions of chromatin structure known as DNaseI hypersensitive sites (DHSs), because these sites are a prerequisite for any type of *cis*-regulatory activity, including enhancers, silencers, insulators, and boundary elements. We demonstrated that DHSs exhibit a distinct sequence signature, which can be used to predict with high accuracy hypersensitive locations in the human genome [61]. We used these signatures to predict novel hypersensitive sites, which were then validated via qPCR and Southern blot analysis. Subsequently, we demonstrated in a series of papers that the converse phenomenon, well-positioned nucleosomes, can be predicted with high accuracy [67, 33, 74, 73]. At the same time, we collaborated with several research groups in the development of high-throughput assays for interrogating local chromatin structure in the human genome [76, 21]. And we designed computational methods capable of identifying, from high-resolution DNaseI sequencing data, all of the DNA-binding footprints in a given genome [35, 14].

Our work on chromatin structure has been carried out within the context of the ENCODE consortium [25]. During the first phase of the project, we developed tools to integrate data on DNaseI sensitivity, replication timing, histone modifications, bulk RNA transcription, and regulatory factor binding region density. In particular, we combined wavelet analyses and hidden Markov models [19] to simultaneously visualize and segment multiple genomic data

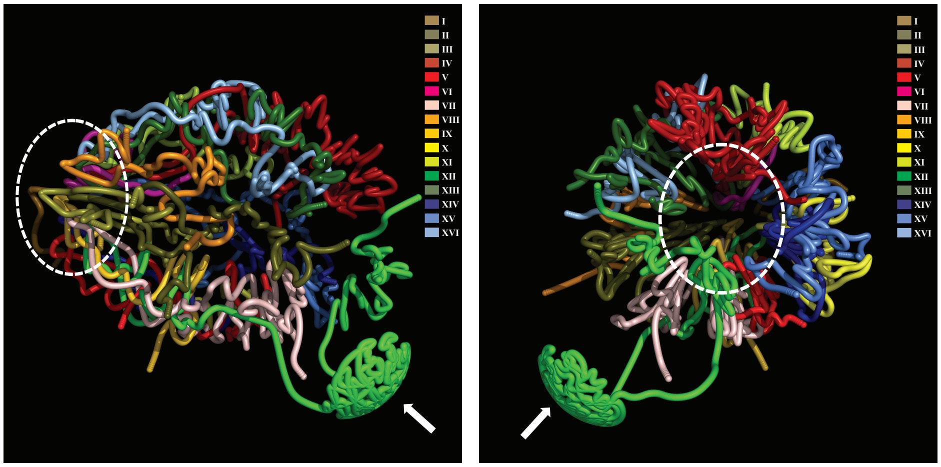

Figure 3: **Three-dimensional model of the yeast genome.** Two views representing two different angles are provided. Chromosomes are coloured as indicated in the upper right. All chromosomes cluster via centromeres at one pole of the nucleus (the area within the dashed oval), while chromosome XII extends outward towards the nucleolus, which is occupied by rDNA repeats (indicated by the white arrow). After exiting the nucleolus, the remainder of chromosome XII interacts with the long arm of chromosome IV.

sets at a variety of scales. The results of these analyses were reported in the ENCODE paper [26] (see Figure 2), as well as in a companion paper [84]. During the current, second phase of ENCODE, my lab is funded as part of the ENCODE Data Analysis Center, and I lead the "Large-scale behavior" analysis group, which focuses on developing methods to perform joint unsupervised learning on multiple tracks of results from sequence census assays such as chromatin immunoprecipitation-sequencing (ChIP-seq) or DNase-seq. Toward this end, we have developed a DBN software system capable of jointly analyzing dozens of parallel tracks of genomic data at base-pair resolution. The resulting model allows us to identify multiple levels of chromatin organization and the functional elements therein [**?**].

Most recently, in collaboration with Tony Blau, Stan Fields and Jay Shendure, we developed a novel method to globally capture intra- and inter-chromosomal interactions, and applied it to generate a map at kilobase resolution of the haploid genome of *Saccharomyces cerevisiae* [**?**]. The map recapitulates known features of genome organization, thereby validating the method, and identifies new features. Extensive regional and higher order folding of individual chromosomes is observed. Chromosome XII exhibits a striking conformation that implicates the nucleolus as a formidable barrier to interaction between DNA sequences at either end. Inter-chromosomal contacts are anchored by centromeres and include interactions among transfer RNA genes, among origins of early DNA replication and among sites where chromosomal breakpoints occur. Finally, we constructed a three-dimensional model of the yeast genome. Our findings provide a glimpse of the interface between the form and function of a eukaryotic genome. For this paper, of which Tony Blau and I are co-corresponding authors, the assay development was carried out by a postdoc in Tony's lab, Stan Fields provided expertise related to yeast, Jay Shendure provided the sequencing technology, and my lab developed methods for assigning statistical confidence measures to the observed interactions, a variety of techniques for relating the observed interactions to known functional elements, and an optimization framework for inferring the 3D model.

In the future, my research in this area will follow three complementary threads. First, we will develop and apply algorithms for characterizing the motif composition of DHSs. My PhD research focused on algorithms for identifying and searching with protein and DNA sequence motifs [31, 29, 30, 4], and I have continued to work in this area, developing new statistical methods for searching for *cis*-regulatory modules [3] and for quantifying similarity between motifs [34]. We have used our methods to identify a yeast transcription factor (Hcm1) that fills the S phase gap in the transcriptional circuitry of the cell [69]. We expect DHSs to be significantly enriched for transcription factor binding sites; therefore, we will search our growing library of DHSs, using known motifs as well as *de novo* motif discovery algorithms and taking into account the observed degree of evolutionary conservation, as well as the accompanying patterns of histone modifications. In any single tissue, only a small portion of observed DHSs are constitutively active. Hence, we are particularly interested in segregating the DHSs according to their tissue specificity, and according to the mRNA expression profiles of their proximal genes, thereby identifying motifs that are tissue- or condition-specific.

Second, we will develop methods that identify and classify functional elements. Our Segway system allows us to identify functional elements using *semi-supervised learning*, in which a small collection of known functional elements is provided to the system, along with a large set of unlabeled data. The system looks for joint patterns across a range of given data sets and can automatically identify novel patterns not associated with any known label, as well as significant subcategories of known labels. Using this method, for example, we can "rediscover" protein-coding genes purely on the basis of histone modification and TF binding data. We plan to extend this approach to identify and decode the complex patterns of histone modifications associated with various types of functional elements—insulators, silencers, enhancers, active and inactive promoters. Also, in an ongoing NSF-funded collaboration with Zhiping Weng's lab at UMass, we will characterize sequence patterns associated with well-positioned nucleosomes, and we will investigate the evolution of these patterns among yeast, fly, mouse and human.

Third, we will continue to investigate the large-scale properties of chromatin structure, specifically in relation to the 3D model described above. With Tony Blau's lab, we will apply the 3D modeling approach to the human genome. We will then compare the resulting model with the large-scale chromatin structure inferred via Segway on various human cell lines. Ultimately, we hope to more fully understand the relationship between chromatin structure and the large-scale 3D structure of the genome, as well as the implications of these phenomena with respect to gene expression, DNA repair and DNA mutation.

Analysis of mass spectrometry data

Mass spectrometry promises to enable scientists to identify and quantify the entire complement of molecules that comprise a complex biological sample. In biomedicine, mass spectrometry is commonly used in a high-throughput fashion to identify proteins in a mixture. However, the primary bottleneck in this type of experiment is computational. Existing

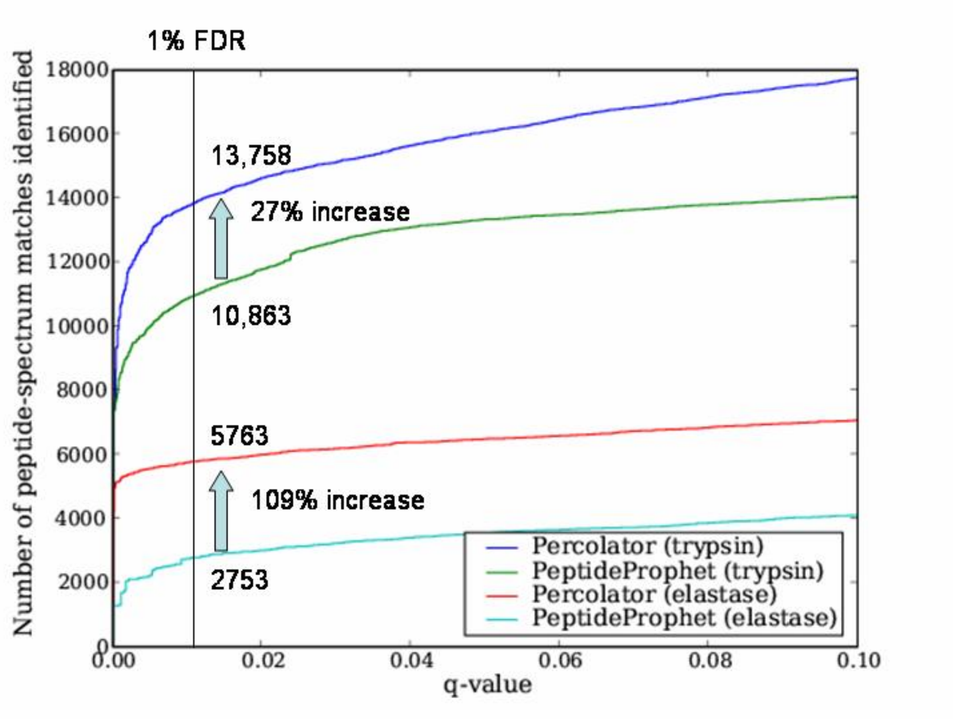

Figure 4: **Comparison of mass spectrum peptide identification methods.** The figure plots the number of spectra identified, as a function of false discovery rate, for two data sets and two analysis methods. For typical data, digested with the standard enzyme trypsin, Percolator improves the identification rate by 27% at a 1% false discovery rate. When we switch to a non-standard enzyme, elastase, Percolator yields more than twice as many identifications.

algorithms for interpreting mass spectra are slow and fail to identify a large proportion of the given spectra.

My lab has made at least four significant contributions to the field of mass spectrometry analysis. The first contribution was to introduce, in 2001, the idea of applying machine learning methods to this type of data [1]. This work was carried out in parallel with similar work at the Institute for Systems Biology [42]. Subsequently, many other groups have applied machine learning methods to mass spectrometry analysis [24, 53, 68, 87, 28].

One significant challenge in applying machine learning to mass spectra is the variability of the data due to different types of samples (*e.g.*, soluble versus membrane proteins), enzyme specificity, modified versus unmodified peptides, mass spectrometer type, database size, instrument calibration, etc. Our second contribution to the field of mass spectometry has been to solve this problem by applying a technique known as semi-supervised learning to the classification of peptide-spectrum matches (PSMs) [37]. In semi-supervised learning, the training set consists of two subsets of examples, one subset with labels and one without. In this application, we search a given set of spectra against two databases, the real (“target”) database and a shuffled (“decoy”) version of the same database. PSMs against the decoy database can be confidently labeled as incorrect identifications, but PSMs against the target database are comprised of a mixture of correct and incorrect identifications. We designed an iterative, semi-supervised algorithm in which the inner loop is an SVM classifier. The algorithm, called Percolator, can be applied to any given mass spectrometry data set, learning model parameters that are appropriate for those data. Relative to a state-of-the-art fully supervised machine learning method, this semi-supervised approach more than doubles the number of correctly identified peptides for some data sets (see Figure 4). A follow-up paper written by a different research group described how to adapt Percolator to the Mascot search engine [10]. Percolator is now being distributed free along with every copy of Mascot, (`http://www.matrixscience.com/pdf/2009WKSHP5.pdf`) which is the most widely used proteomics search engine. Also, Thermo's Differential analysis software, called Sieve (`http://www.thermo.com/com/cda/product/detail/1,1055,10123438,00.html`) uses Percolator to improve the ability of SEQUEST to identify the differences that it finds. The popular PeptideProphet software [42] was subsequently updated to include a semi-supervised learning mode [16].

In the field of mass spectrometry analysis, the proper definition and application of methods for estimating statistical confidence measures is the subject of ongoing debate [40, 32, 15, 27, 56]. Our third contribution has been to describe how to compute rigorous statistical confidence measures. For example, we have described empirical methods for calibrating an existing score function—the SEQUEST XCorr [43]. And in collaboration with John Storey, we have described how to apply existing methods from the statistical literature [8, 82, 83] to mass spectrometry data [39, 40, 38, 41], emphasizing the need for multiple testing correction via false discovery rate analysis [60].

Finally, our fourth contribution has been to make the field of mass spectrometry more open. When I first started publishing in this field, it was rare for research groups to make their primary data publicly available. Indeed, it was possible to publish a paper simply reporting on the availability of a new data set [70]. The source code for many widely used software packages, such as SEQUEST and Mascot, was not available. And the SEQUEST patent, held by the University of Washington, was demonstrably hindering development of new techniques by discouraging researchers from working in this area. I have publicly criticized the dearth of publicly available data [57], and I have consistently made freely available the benchmark data sets used in our studies. Furthermore, I have successfully negotiated with the university's tech transfer office to make the license to the SEQUEST patent, held by Thermo Scientific, non-exclusive, and I have then published a reimplementation of SEQUEST which is freely available, with source code, for academic and non-profit users [64]. Finally, we have dramatically expanded on this core SEQUEST functionality, providing a rich software toolkit that provides database search functionality, powerful machine learning algorithms [37, 81], accurate statistical confidence measures [43], and the ability to find post-translational modifications and cross-linked peptides [54]. This toolkit has a growing user community and is under active development. Indeed, we have recently improved the speed of the core search engine by three orders of magnitude [22].

In the future, we plan to employ a collection of cooperating dynamic Bayesian networks to model jointly an entire mass spectrometry experiment. Relative to most existing methods for analyzing mass spectrometry data, which tend to divide the analysis of an experiment into a series of small independent subtasks, this unified model will jointly consider all of the available data. This approach can thus exploit valuable dependencies among spectra and along various dimensions of the data. Dynamic Bayesian networks also provide a rigorous framework for performing inference from a combination of observed data and qualitative expert knowledge.

I have recently laid out this plan, in an R01 competing renewal that was divided into five aims, each of which concerns a particular type of mass spectrometry experiment. These experiments involve (1) identifying all of the proteins in a given complex biological sample using a standard mass spectrometry protocol; (2) identifying proteins using a modified protocol in which the mass spectrometer samples the data in a systematic, rather than data-dependent, fashion, with the goal of identifying lower abundance proteins; (3) quantifying the relative abundance of proteins within or between biological samples; (4) identifying post-translationally modified proteins or proteins that contain sequence variation; and (5) performing targeted quantification of a specified set of proteins, such as proteins in a pathway of interest or protein biomarkers.

Such methods have the potential to dramatically improve our ability to draw conclusions from and formulate hypotheses on the basis of high-throughput shotgun proteomics experiments. Experiments like the ones described above can, for example, identify proteins involved in fundamental disease processes, identify previously unknown protein isoforms, or quantify the responses of proteins to environmental stressors or disease states.

References

[1] D. C. Anderson, W. Li, D. G. Payan, and W. S. Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and Sequest scores. *Journal of Proteome Research*, 2(2):137–146, 2003.

[2] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, C. Cortes, and M. Mohri. Polynomial semantic indexing. In *Advances in Neural Information Processing Systems*, 2009.

[3] T. L. Bailey and W. S. Noble. Searching for statistically significant regulatory modules. *Bioinformatics*, 19(Suppl. 2):ii16–ii25, 2003.

[4] M. E. Baker, W. N. Grundy, and C. P. Elkan. Spinach CSP41, an mRNA-binding protein and ribonuclease, is homologous to nucleotide-sugar epimerases and hydroxysteroid dehydrogenases. *Biochemical and Biophysical Research Communications*, 248(2):250–254, 1998.

[5] A. Ben-Hur and D. Brutlag. Remote homology detection: a motif based approach. *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, 19 suppl 1:i26–i33, 2003.

[6] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 suppl 1:i38–i46, 2005.

[7] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2000.

[8] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

[10] M. Brosch, L. Yu, T. Hubbard, and J. Choudhary. Accurate and sensitive peptide identification with Mascot Percolator. *Journal of Proteome Research*, 8(6):3176–3181, 2009.

[11] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262–267, 2000.

[12] S. Busuttil, J. Abela, and G. J. Pace. Support vector machines with profile-based kernels for remote protein homology detection. *Genome Informatics*, 15(2):191–200, 2004.

[13] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-supervised Learning.* MIT Press, Cambridge, MA, 2006.

[14] X. Chen, M. M. Hoffman, J. A. Bilmes, J. R. Hesselberth, and W. S. Noble. A dynamic Bayesian network for identifying protein binding footprints from single molecule based sequencing data. *Bioinformatics*, 26(12):i334–i342, 2010.

[15] H. Choi and A. I. Nesvizhskii. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(1):47–50, 2007.

[16] H. Choi and A. I. Nesvizhskii. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(1):254–265, 2008.

[17] R. Collobert and J. Weston. Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 2007.

[18] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

[19] N. Day, A. Hemmaplardh, R. E. Thurman, J. A. Stamatoyannopoulos, and W. S. Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007.

[20] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 140–151, 2003.

[21] J. H. Dennis, H. Fan, S. M. Reynolds, G. Yuan, J. C. Meldrim, D. J. Richter, D. G. Peterson, O. J. Rando, W. S. Noble, and R. E. Kingston. Independent and complementary methods for large-scale structural analysis of mammalian chromatin. *Genome Research*, 17(6):928–939, 2007.

[22] B. Diament and W. S. Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.

[23] Q. W. Dong, X. L. Wang, and L. Lin. Application of latent semantic analysis to protein remote homology detection. *Bioinformatics*, 22(3):285–290, 2006.

[24] J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth, and S. P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22:214–219, 2004.

[25] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306(5696):636–640, 2004.

[26] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[27] M. Fitzgibbon, Q. Li, and M. McIntosh. Modes of inference for evaluating the confidence of peptide identifications. *Journal of Proteome Research*, 7(1):35–39, 2008.

[28] A. M. Frank. A ranking-based scoring function for peptide-spectrum matches. *Journal of Proteome Research*, 8(5):2241–2252, 2009.

[29] W. N. Grundy. *A Bayesian Approach to Motif-based Protein Modeling.* PhD thesis, University of California, San Diego, La Jolla, CA, 1998.

[30] W. N. Grundy, T. L. Bailey, C. P. Elkan, and M. E. Baker. Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochemical and Biophysical Research Communications*, 231(3):760–766, 1997.

[31] W. N. Grundy, T. L. Bailey, C. P. Elkan, and M. E. Baker. Meta-MEME: motif-based hidden Markov models of protein families. *Computer Applications in the Biosciences*, 13(4):397–406, 1997.

[32] N. Gupta and P. Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of Proteome Research*, 8(9):4173–4181, 2009.

[33] S. Gupta, J. Dennis, R. E. Thurman, R. Kingston, J. A. Stamatoyannpoulos, and W. S. Noble. Predicting human nucleosome occupancy from primary sequence. *PLOS Computational Biology*, 4(8):e10000134, 2008.

[34] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble. Quantifying similarity between motifs. *Genome Biology*, 8:R24, 2007.

[35] J. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods*, 6(4):283–289, 2009.

[36] E. Ie, J. Weston, W.S. Noble, and C. Leslie. Adaptive codes for multi-class protein classification. In *Proceedings of the International Conference on Machine Learning*, 2005.

[37] L. Käll, J. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–25, 2007.

[38] L. Käll, J. Storey, and W. S. Noble. Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*, 24(16):i42–i48, 2008.

[39] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(1):29–34, 2008.

[40] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of Proteome Research*, 7(1):40–44, 2008.

[41] L. Käll, J. D. Storey, and W. S. Noble. QVALITY: Nonparametric estimation of q values and posterior error probabilities. *Bioinformatics*, 25(7):964–966, 2009.

[42] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Analytical Chemistry*, 74:5383–5392, 2002.

[43] A. A. Klammer, C. Y. Park, and W. S. Noble. Statistical calibration of the SEQUEST XCorr function. *Journal of Proteome Research*, 8(4):2106–2113, 2009.

[44] A. A. Klammer, S. R. Reynolds, M. Hoopmann, M. J. MacCoss, J. Bilmes, and W. S. Noble. Modeling peptide fragmentation with dynamic Bayesian networks yields improved tandem mass spectrum identification. *Bioinformatics*, 24(13):i348–i356, 2008.

[45] A. A. Klammer, C. C. Wu, M. J. MacCoss, and W. S. Noble. Peptide charge state determination for low-resolution tandem mass spectra. In *Proceedings of the Computational Systems Bioinformatics Conference*, pages 175–185, 2005.

[46] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. *Journal of Bioinformatics and Computational Biology*, 3(3):527–550, 2005.

[47] P. Kuksa, P. Huang, and V. Pavlovic. Fast and accurate multi-class protein fold recognition with spatial sample kernels. In *Computational Systems Bioinformatics: Proceedings of the CSB2008 Conference*, pages 133–143, 2008.

[48] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[49] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 300–311. World Scientific, 2004.

[50] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 564–575, New Jersey, 2002. World Scientific.

[51] C. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classification. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 1441–1448, Cambridge, MA, 2003. MIT Press.

[52] L. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology*, pages 225–232, Washington, DC, April 18–21 2002.

[53] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*, 25:125–131, 2006.

[54] S. McIlwain, P. Draghicescu, P. Singh, D. R. Goodlett, and W. S. Noble. Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. *Journal of Proteome Research*, 9(5):2488–2495, 2010.

[55] I. Melvin, E. Ie, R. Kuang, J. Weston, W. S. Noble, and C. Leslie. SVM-fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics*, 8(Suppl 4):S2, 2007.

[56] P. Navarro and J. Vazquez. A refined method to calculate false discovery rates for peptide identification using decoy databases. *Journal of Proteome Research*, 8(4):1792–1796, 2009.

[57] W. S. Noble. Data hoarding is harming proteomics. *Nature Biotechnology*, 22:1209, 2004.

[58] W. S. Noble. Support vector machine applications in computational biology. In B. Schoelkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel methods in computational biology*, pages 71–92. MIT Press, Cambridge, MA, 2004.

[59] W. S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

[60] W. S. Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, 2009.

[61] W. S. Noble, S. Kuehn, R. Thurman, R. Humbert, J. C. Wallace, M. Yu, M. Hawrylycz, and J. Stamatoayannopoulos. Predicting the *in vivo* signature of human gene regulatory sequences. *Bioinformatics*, 21(Suppl 1):i338–i343, 2005.

[62] H. Ogul and E. U. Mumcuoglu. SVM-based detection of distant protein structural relationships using pairwise probabilistic suffix trees. *Computational and Biological Chemistry*, 30(4):292–299, 2006.

[63] H. Ogul and E. U. Mumcuoglu. A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabet. *Biosystems*, 87(1):75–81, 2007.

[64] C. Y. Park, A. A. Klammer, L. Käll, M. P. MacCoss, and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research*, 7(7):3022–3027, 2008.

[65] P. Pavlidis, T. S. Furey, M. Liberto, D. Haussler, and W. N. Grundy. Promoter region-based classification of genes. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing 2001*, pages 151–163, Singapore, 2001. World Scientific.

[66] P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology*, pages 242–248, 2001.

[67] H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. Nucleosome positioning signals in genomic DNA. *Genome Research*, 17(8):1170–1177, 2007.

[68] K. Petritis, L. J. Kangas, P. L. Ferguson, G. A. Anderson, L. Pasa-Tolic, M. S. Lipton, K. J. Auberry, E. F. Strittmatter, Y. Shen, R. Zhao, and R. D. Smith. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Analytical Chemistry*, 75(5):1039–1048, 2003.

[69] T. Pramila, W. Wu, W. S. Noble, and L. L. Breeden. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S phase gap in the transcriptional circuitry of the cell cycle. *Genes and Development*, 20(16):2266–2278, 2006.

[70] S. Purvine, A. F. Picone, and E. Kolker. Standard mixtures for proteome studies. *OMICS*, 8:79–92, 2004.

[71] J. Qiu and W. S. Noble. Predicting co-complexed protein pairs from heterogeneous data. *PLOS Computational Biology*, 4(4):e1000054, 2008.

[72] H. Rangwala and G. Karypis. Profile based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, 21:4239–4247, 2005.

[73] S. Reynolds, Z. Weng, J. Bilmes, and W. S. Noble. Predicting nucleosome positioning using multiple evidence tracks. In *Proceedings of the Fourteenth Annual International Conference on Computational Molecular Biology*, volume 6044/2010, pages 441–445, 2010.

[74] S. M. Reynolds, J. Bilmes, and W. S. Noble. Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in *Saccharomyces cerevisiae* and *Homo sapiens*. *PLOS Computational Biology*, 6(7):e10000834, 2010.

[75] S. M. Reynolds, L. Käll, M. E. Riffle, J. A. Bilmes, and W. S. Noble. Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. *PLOS Computational Biology*, 4:e1000213, 2008.

[76] P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Genome-wide identification of DNase1 hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4537–4542, 2004.

[77] H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

[78] B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

[79] N. H. Segal, P. Pavlidis, C. R. Antonescu, R. G. Maki, W. S. Noble, J. M. Woodruff, J. J. Lewis, M. F. Brennan, A. N. Houghton, and C. Cordon-Cardo. Classification and subtype prediction of soft tissue sarcoma by functional genomics and support vector machine analysis. *American Journal of Pathology*, 169:691–700, 2003.

[80] N. H. Segal, P. Pavlidis, W. S. Noble, C. R. Antonescu, A. Viale, U. V. Wesley, K. Busam, H. Gallardo, D. DeSantis, M. F. Brennan, C. Cordon-Cardo, J. D. Wolchok, and A. N. Houghton. Classification of clear cell sarcoma as melanoma of soft parts by genomic profiling. *Journal of Clinical Oncology*, 21:1775–1781, 2003.

[81] M. Spivak, J. Weston, L. Bottou, L. Käll, and W. S. Noble. Improvements to the Percolator algorithm for peptide identification from shotgun proteomics data sets. *Journal of Proteome Research*, 8(7):3737–3745, 2009.

[82] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64:479–498, 2002.

[83] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, Sep 2005.

[84] R. E. Thurman, N. Day, W. S. Noble, and J. A. Stamatoyannopoulos. Identification of higher-order functional domains in the human ENCODE regions. *Genome Research*, 17:917–927, 2007.

[85] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1998.

[86] J.-P. Vert, R. Thurman, and W. S. Noble. Kernels for gene regulatory regions. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1401–1408, Cambridge, MA, 2006. MIT Press.

[87] B.-J. M. Webb-Robertson, W. R. Cannon, C. S. Oehmen, A. R. Shah, V. G., M. S. Lipton, and K. M. Waters. A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, 24(13):1503–9, 2008.

[88] J. Weston, C. Leslie, D. Zhou, and W. S. Noble. Semi-supervised protein classification using cluster kernels. In *Advances in Neural Information Processing Systems 16*, pages 595–602, 2004.