Vorhersage von DNA-Bindungsstellen mit generativen, diskriminativen und hybriden Lernverfahren

Jens Keilwagen

Abteilung Molekulare Genetik
Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) Gatersleben
Jens.Keilwagen@ipk-gatersleben.de

Abstract: Probabilistische Modelle werden heutzutage aufgrund ihrer Flexibilität in vielen Bereichen zur Modellierung und Klassifikation von anfallenden Daten genutzt. Von entscheidender Bedeutung ist neben der Wahl des entsprechenden Modells auch die Wahl des Lernverfahrens, welches die Modellparameter aus gegebenen Daten inferriert. Häufig wird dieser Aspekt völlig aus den Augen gelassen, obwohl er sehr viel Potenzial birgt. In der vorgelegten Dissertation wird u.a. ein verallgemeinertes Lernverfahren vorgestellt und auf biologische Daten angewendet. Die objektorientierte und quelloffene Implementierung ermöglicht eine Vielzahl weiterer Anwendungen und Erweiterungen.

1 Einführung

"Wir ertrinken in Informationen - aber wir hungern nach Wissen!"
(Rutherford D. Rogers, NY Times, 1988)

Im fortschreitenden Informationszeitalter mit der Erhebung von Massendaten auf allen Ebenen der Wirtschaft, der Wissenschaft und des Lebens ist Rutherford D. Rogers Aussage lebendiger und greifbarer denn je. Im Zeitalter großer Suchmaschinen, in denen Millionen an Informationen nur einen Klick entfernt sind und historische als auch aktuelle Daten neue Möglichkeiten eröffnen, bedarf es geeigneter, computergestützter Systeme, um dem Nutzer ein Maximum an sinnvollen Informationen zu präsentieren und ihm damit den Weg zum Wissen zu ebnen.

Gerade in den Wissenschaften stieg in den letzten Jahren der Bedarf der Datenauswertungen durch die Entwicklung neuer Hochdurchsatzexperimente rasant an. Neben der Medizin gehören vor allem die experimentellen Wissenschaften, wie Chemie, Physik und Biologie zu den großen Datenproduzenten. Dabei ist neben der Speicherung und Visualisierung dieser Daten besonders die Klassifikation und die Hypothesenbildung in den Fokus der Forschung gerückt. Diese Aufgaben prägen das Forschungsgebiet des maschinellen Lernens, dessen Vorhersagen einen entscheidenden Einfluss auf die weiteren Forschungsrichtungen anderer Disziplinen haben.

Im Bereich der molekularen Genetik sind dabei vor allem Analysen von Bindungs-

stellen auf der DNA und ihrer Arbeitskopie der mRNA, die ein wichtiger Bestandteil in der Regulation der Aktivität von Genen sind, in den Fokus des maschinellen Lernen gerückt. Diese Analysen umfassen zum Beispiel die Erkennung von Transkriptionsfaktorbindungsstellen [KGR+03, BEFK03], Transkriptionstartstellen [SZR06, AVS09], Donor- und Akzeptorspleißstellen [BK97, Sal97, YB04], Nukleosomenbindungsstellen [SFMC+06, PTF+07], miRNA-Bindungsstellen [LhSJR+03, MRS+09], und Insulatorbindungsstellen [KAS+07].

In den Analysen werden dabei häufig probabilistische Modelle genutzt, um die gegebenen Daten zu modellieren und später eine Klassifikation neuer Daten zu ermöglichen. Probabilistische Modelle definieren dabei eine Menge von Annahmen, die sich als bedingte statistische Unabhängigkeiten schreiben lassen. Die Leistungsfähigkeit eines probabilitischen Modells hängt dabei jedoch nicht nur von diesen Annahmen sondern auch von den Modellparametern ab, die aus den gegebenen Daten gelernt werden.

Für die Inferenz der Modellparameter werden unterschiedliche Lernverfahren genutzt, die in verschiedene Kategorien aufgeteilt werden können. Lernverfahren, die die Modellparameter ausschließlich aus den gegebenen Daten bestimmen, nennt man *nicht-bayesisch*, während man solche, die auch Vorwissen in Form einer A-Priori-Verteilung in die Bestimmung der Modellparameter eingehen lassen, als *bayesisch* bezeichnet.

Darüber hinaus kann man Lernverfahren auch anhand ihrer Zielfunktion unterscheiden. *Generative* Lernverfahren zielen auf die exakte Repräsentation der Verteilung der Trainingsdaten, während *diskriminative* Lernverfahren auf eine exakte Klassifikation der Trainingsdaten zielen. Neben diesen beiden sich konträr gegenüberstehenden Lernverfahren gibt es eine Reihe von *hybriden* Lernverfahren, die eine Interpolation zwischen den beiden Zuvorerwähnten ermöglichen.

In der vorgelegten Dissertation wurde ein vereinheitlichtes Lernverfahren vorgestellt, das es ermöglicht, zwischen sechs etablierten Lernverfahren zu interpolieren – unter ihnen bayesische, nicht-bayesische, generative, diskriminative und hybride. Für Anwendung dieses Lernverfahrens auf Markov Random Fields wurde zudem eine A-Priori-Verteilung hergeleitet, die in Spezialfällen eine bekannte A-Priori-Verteilung einfacherer Modelle ist. Auf Basis dieser theoretischen Arbeiten können nun probabilistische Modelle effizienter für die Klassifikation unterschiedlicher Daten einschließlich biologischer Sequenzdaten genutzt werden.

2 Parameterlernen probabilistischer Modelle

Probabilistische Modelle beschreiben eine Verteilung für einen Ereignisraum, in dem sie jedem Ereignis eine Wahrscheinlichkeit zuordnen. Betrachten wir den Ereignisraum aller Sequenzen $\underline{x}$ mit Label c, dann beschreibt ein solches Modell die Verteilung $P(c, \underline{x}|\underline{\lambda})$, wobei die Modellparameter $\underline{\lambda}$ die Verteilung und damit die Wahrscheinlichkeit für jedes Paar $(c, \underline{x})$ spezifizieren.

2.1 Lernverfahren

Für die Klassifikation neuer Daten ist die Wahl der Modellparameter und damit die Frage *"Wie erhält man* gute *Modellparameter?"* entscheidend. In den letzten Jahren gab es eine Reihe von Neu- und Weiterentwicklungen auf diesem Forschungsgebiet. Im weiteren beschränken wir uns auf sechs weit verbreitete Lernverfahren.

Vergleicht man die Zielfunktionen des nicht-bayesischen, generativen *Maximum Likelihood* (ML) Prinzips [Fis22], des bayesischen, generativen *Maximum A-Posteriori* (MAP) Prinzips [Bis06], des nicht-bayesischen, diskriminativen *Maximum Conditional Likelihood* (MCL) Prinzips [WGR$^+$02], des bayesischen, diskriminativen *Maximum Supervised Posteriori* (MSP) Prinzips [GKM$^+$02], des nicht-bayesischen, hybriden *Generative-Discriminative Trade-Off* (GDT) Prinzips [BT04] und des bayesischen, hybriden *Penalized Generative-Discriminative Trade-Off* (PGDT) Prinzips [BT04], so stellt man fest, dass drei Terme ausreichen, um diese Lernverfahren zu definieren:

1. die *Conditional Likelihood* $P(c|\underline{x},\underline{\lambda}) = \frac{P(c,\underline{x}|\underline{\lambda})}{\sum_{\tilde{c}} P(\tilde{c},\underline{x}|\underline{\lambda})}$,
2. die *Likelihood* $P(c,\underline{x}|\underline{\lambda})$, und
3. die A-Priori-Verteilung $Q(\underline{\lambda}|\underline{\alpha})$.

Mit dem Ziel der Vereinigung und Verallgemeinerung dieser sechs Lernverfahren, schlagen wir ein verallgemeinertes generative-diskriminatives Lernverfahren vor, das die Modellparameter $\underline{\lambda}$ als Maximum des gewichteten Produkts der Terme Conditional Likelihood, Likelihood, and A-Priori-Verteilung bestimmt [**?**]. Die Zielfunktion des Lernverfahren lautet für einen Datensatz mit N Sequenzen

$$\hat{\underline{\lambda}} = \underset{\underline{\lambda}}{\operatorname{argmax}} \left[\beta_0 \left[\sum_{n=1}^{N} \log P(c_n|\underline{x}_n,\underline{\lambda}) \right] + \beta_1 \left[\sum_{n=1}^{N} \log P(c_n,\underline{x}_n|\underline{\lambda}) \right] + \beta_2 \log Q(\underline{\lambda}|\underline{\alpha}) \right]$$

mit den Wichtungsfaktoren $\underline{\beta} := (\beta_0,\beta_1,\beta_2)$, $\beta_0,\beta_1,\beta_2 \in [0,1]$, und $\beta_0+\beta_1+\beta_2 = 1$. Wir erhalten die sechs etablierten Lernverfahren als Spezialfälle wie in Abbildung 1 dargestellt.

2.2 A-Priori Verteilung für Markov Random Fields

Die vorgestellten Lernverfahren lassen sich grundsätzlich auf alle probabilitischen Modelle anwenden. Eine Klasse von probabilitischen Modellen, die sehr flexibel ist und sich daher zum Beispiel hervorragend für die Klassifikation von Spleißstellen eignet, sind Markov Random Fields [YB04]. Viele einfachere Modelle wie inhomogene [Sal97] und permutierte [EYSJ02] Markov Modelle gehören zu der Klasse der Markov Random Fields. Während es für diese Modelle eine gut interpretierbare und weit verbreitete A-Priori-Verteilung gibt, fehlte eine vergleichbare Verteilung für Markov Random Fields.

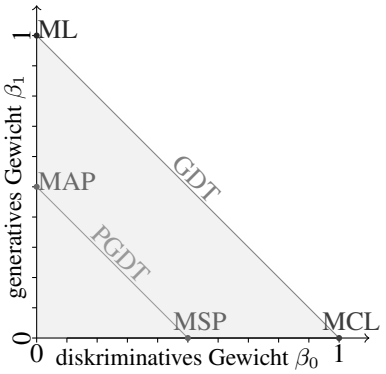


Abbildung 1: Zweidimensionale Projektion des durch die Wichtungsfaktoren $\underline{\beta}$ aufgespannten Simplexes des verallgemeinerten generative-diskriminativen Lernverfahrens. Die Punkte $(0, 1)$, $(0, 0.5)$, $(1, 0)$, and $(0.5, 0)$ korrespondieren mit den ML, MAP, MCL, und MSP Prinzipien, während die Geraden $\beta_1 = 1 - \beta_0$ und $\beta_1 = 0.5 - \beta_0$ dem GDT und dem PGDT Prinzip entsprechen.

Um die Ergebnisse von Markov Random Fields mit denen von einfacheren Modellen vergleichbar zu machen, nutzen wir das gleiche Lernverfahren. Dies war im Falle von bayesischen Lernverfahren so nicht möglich, da eine vergleichbare und allgemeine A-Priori-Verteilung für Markov Random Fields fehlte. Die Herleitung einer entsprechenden A-Priori-Verteilung [**?**] ermöglicht eine verbesserte Analyse der Modellannahme und erlaubt damit fundierte Aussagen über die zu analysierenden Daten.

Abbildung 2 zeigt die A-Priori-Verteilung für einen und zwei freie Parameter. Untersucht man die hergeleitetete Verteilung für einen freien Parameter, so stellen wir fest, dass sie sich am Maximum ähnlich einer Gauß- und in den Schwänzen ähnlich einer Laplaceverteilung verhält. Für zwei freie Parameter erhalten wir bereits eine unsymmetrische Verteilung, die weder mit einer mehrdimensionalen Gauß- noch mit einer mehrdimensionalen Laplaceverteilung vergleichbar ist.

Um eine zügige Analyse von Sequenzdaten zu ermöglichen, haben wir die Komponenten probabilitische Modelle, Lernverfahren und A-Priori-Verteilung objektorientiert in der quelloffenen Softwarebibliothek Jstacs [KGG$^+$08] implementiert. Dadurch können zum einen praktische Problemstellungen wie Klassifikations- oder Modellierungsaufgaben rasch umgesetzt werden. Andererseits erlaubt es das schnelle Implementieren und Testen neuer Komponenten, seien es Modelle, Lernverfahren oder A-Priori-Verteilungen.

3 Probabilistische Modelle in der Bioinformatik

Nachdem wir im letzten Kapitel vor allem theoretische Aspekte angesprochen haben, werden wir uns in diesem Kapitel drei praktischen Anwendungen probabilistischer Modelle

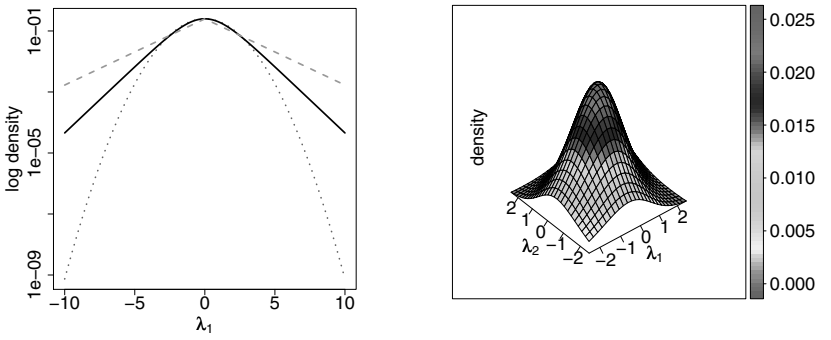


Abbildung 2: Visualisierung der A-Priori-Verteilung für einen und zwei freie Parameter. In linken Abbildung werden neben der hergeleiteten Verteilung auch der Gauß- und Laplaceverteilung gepunktet bzw. gestrichelt dargestellt.

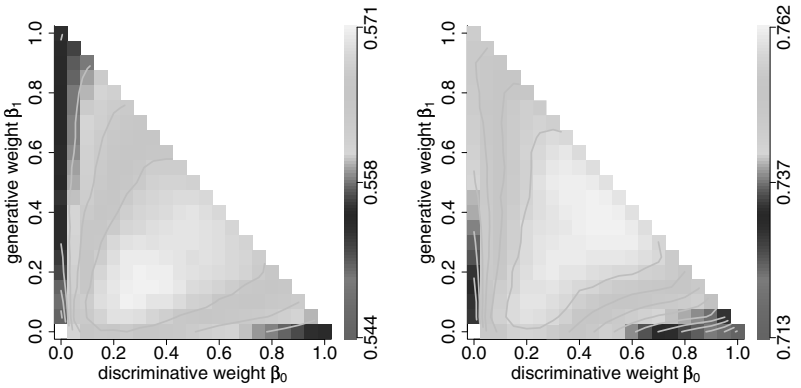
in der Bioinformatik widmen.

3.1 Transkriptionsfaktorbindungsstellenerkennung

Die Vorhersage von Transkriptionsfaktorbindungsstellen basierend auf einer Menge ähnlicher Bindungsstellen ist eine zentrale Aufgabe, um unser Verständnis der Genregulation zu verbessern. Zur Lösung dieser Aufgabe werden Modelle unter Verwendung bereits publizierter Bindungsstellen gebildet, die anschließend zur Klassifikation und damit Vorhersage neuer Bindungsstellen genutzt werden können.

Traditionell wird dabei auf generativ gelernte probabilistische Modelle zurückgegriffen. Basierend auf den Ausführungen im letzten Kapitel, können wir das generative Lernverfahren durch das verallgemeinerte generative-diskriminative Lernverfahren austauschen [**?**]. Abbildung 3 zeigt die gemittelten Ergebnisse eines 1.000-fachen Hold-out Experiments für die Transkriptionsfaktorenfamilien AR/GR/PR und NF-κB projeziert auf den Simplex der Wichtungsfaktoren $\underline{\beta}$ (vgl. Abb. 1). Wir finden die besten Ergebnisse im Innerern des Simplexes, was darauf hindeutet, dass ein hybrides Lernverfahren am besten funktioniert.

Vergleichen wir das jeweils beste Ergebnis mit den entsprechenden Ergebnissen für die traditionell verwendeten ML und MAP Prinzipien, dann finden wir eine signifikante Verbesserung der Güte. Diese Beobachtung deutet darauf hin, dass neben der Wahl des Modells auch die Wahl des Lernverfahren einen entscheidenden Einfluss auf die Güte der Vorhersage hat.



AR/GR/PR NF-κB

Abbildung 3: Gemittelte auc-PR für die Transkriptionsfaktorbindungsstellendatensatz AR/GR/PR und NF-κB basierend auf einem 1.000-fachen Hold-out Experiment unter Nutzung des verallgemeinerten generative-diskriminative Lernverfahrens. Helle Bereiche entlang der Winkelhalbierenden zeigen gute Resultate, während dunkle Bereiche auf den Achsen weniger gute Resultate visualisieren. Die Konturlinien symbolisieren Vielfache des Standardfehlers des besten Ergebnisses.

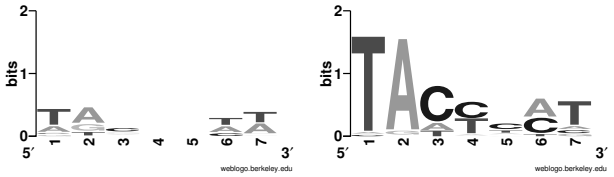
3.2 Erkennung von Annotationsfehlern in biologischen Datenbanken

Aufgrund der rasant zunehmenden Zahl von biologischen Experimenten und dem damit verbundenen Erkenntnisgewinn ist die persistente Speicherung der Ergebnisse nicht mehr wegzudenken. Insbesondere die Speicherung und Nutzung von annotierten Transkriptionsfaktorbindungsstellen hat in den letzten Jahren einen neuen Markt für IT-Firmen eröffnet [WDKK96].

Datenbanken wie TRANSFAC beziehen dabei ihren Inhalt aus Veröffentlichungen anderer Wissenschaftler. Sogenannte Annotatoren lesen die entsprechenden Veröffentlichungen und fügen die Transkriptionsfaktorbindungsstellen in die Datenbanken ein. Dieser Vorgang birgt Risiken, da sich an unterschiedlichen Stellen Fehler einschleichen können, zum einen während der Veröffentlichung und zum anderen während der Integration in die Datenbanken. Typische Annotationsfehler sind Verschiebungen von Bindungsstellen, falsche oder fehlende Annotation des Doppelstrangs und zusätzliche Annotation von Bindungsstellen.

Zur Identifikation solcher Annotationsfehler analysierten wir die sieben größten Datensätze der frei verfügbaren Datenbank CoryneRegNet [BWKT09] unter Nutzung eines generativ gelernten probabilistischen Modells, das wir MotifAdjuster nennen [**?**].

Wir fanden in diesen sieben Datensätzen, dass mehr als ein Drittel der Daten für eine bedeutende Korrektur vorgeschlagen wurden. Vergleicht man die Sequenzlogos der sieben Datensätze vor und nach der von MotifAdjuster vorgeschlagenen Korrektur, so finden wir



(a) Original (b) Korrigiert

Abbildung 4: Sequenzlogos für Bindungsstellen des Transkriptionsfaktors NarL. Die Abbildung (a) zeigt das Sequenzlogo basierend auf den annotierten Bindestellen der Datenbank, während Abbildung (b) das Sequenzlogo nach der Korrektur durch MotifAdjuster zeigt. Die Größe der Nukleotide symbolisiert den Grad der Konserviertheit. Wir finden daher eine deutlich bessere Übereinstimmung des rechten Sequenzlogos mit der in der Literatur erwähnten Konsensussequenz `TACYYMT`.

eine deutliche bessere Übereinstimmung mit den aus der Literatur bekannten Konsensussequenzen. Beispielhaft zeigt die Abbildung 4 die Sequenzlogos der Bindungsstellen des Transkriptionsfaktors NarL vor und nach der Korrektur.

Bei genauerer Untersuchung der vorgeschlagenen Korrekturen für den NarL Datensatz, finden wir zwei Bindungsstellen, die laut MotifAdjuster entfernt werden sollten, und elf, die verschoben werden sollten. Vergleicht man diese Vorhersagen mit den Angaben in der Originalliteratur, um die Güte von MotifAdjuster zu bestimmen, dann finden wir in zwölf von dreizehn der fraglichen Bindungsstellen eine Bestätigung der Vorhersage. Nur in einem Fall unterscheidet sich die Vorhersage von der Angabe in der Literatur. Bei eingehendem Studium der Literatur stellt man fest, dass diese Bindungsstelle noch kontrovers diskutiert wird.

Zusammenfassend stellen wir fest, dass die computergestützte Vorhersage von Annotationsfehlern in Datenbanken mittels generativ gelernter probabilistischer Modelle sinnvoll ist und tatsächlich auftretende Fehler findet.

3.3 Diskriminative de-novo Motiverkennung

Neben der Vorhersage neuer Bindungsstellen, die bereits annotierten Bindestellen ähneln, ist die de-novo Vorhersage von Bindungsstellen, d.h. die Vorhersage ohne bekannte Beschreibung der Bindungsstellen, in vielen Fragestellungen von großem Interesse. Viele biologische Experimente erlauben es, die Region der Bindung des Proteins an die DNA einzugrenzen, aber die Bindungsstelle nicht exakt zu lokalisieren. Aus diesem Grund werden probabilistische Modelle genutzt, um die Bindungsstellen weiter einzugrenzen.

In den letzten Jahren wurden in diesem Bereich viele neue Verfahren vorgestellt. Zwei Erweiterungen haben sich dabei als besonders vorteilhaft herausgestellt. Zum einen konnte gezeigt werden, dass Programme, die einen diskriminativen Lernverfahren nutzen, häufig

	A-GLAM	DEME	DME	Gibbs Sampler	Improbizer	MEME	Weeder	Dispom
wohlwollend	4	4	0	0	4	2	6	9
strikt	3	0	0	0	2	1	0	9

Tabelle 1: Motiverkennungsresultate für acht Programme und zwei Auswertestrategien. Jede Zeile zeigt die Anzahl der Datensätze, auf denen die Programme jeweils erfolgreich waren.

bessere Resultate liefern als die Programme, die ein generatives Lernverfahren verwenden. Der zusätzliche Kontrolldatensatz, der für die Nutzung eines diskriminativen Lernverfahren notwendig ist, ermöglicht es, zuällig auftretende Teilsequenzen zu ignorieren.

Zum anderen erwies sich die Nutzung einer Positionsverteilung für die Bindungsstellen als sinnvoller zusätzlicher Modellierungsansatz, da Bindungsstellen aufgrund ihrer Funktionalität häufig in der Nähe weiterer Signale auftreten.

Interessanterweise gibt es bisher noch keinen Ansatz der beide Erweiterungen kombiniert. Aus diesem Grund haben wir Dispom, ein diskriminatives Motivsucheprogramm entwickelt, das zusätzlich eine Positionsverteilung der Bindungsstellen aus den Daten lernt [**?**]. Um die Güte dieses Programm mit bereits existierenden Programmen zu vergleichen, haben wir Dispom zusammen mit sieben häufig genutzten Programmen in verschiedenen Szenarien getestet. Diese Programme umfassen Ansätze mit probabilistischen Modellen, die sowohl generativ als auch diskriminativ gelernt werden. Einige diese Programme nutzen sogar eine Positionsverteilung für die Bindungsstellen.

Bei unserem Vergleich stellten wir fest, dass viele Programme unter vereinfachten Annahmen wesentlich besser funktionierten als unter realeren Annahmen. Besonders deutlich zeigt sich der Unterschied zwischen den Programmen, wenn die Länge der Bindungsstellen a-priori unbekannt ist. Tabelle 1 zeigt die Resultate der acht Programme für neun verschiedene Datensätze und zwei verschiedene Auswertungsstrategien. In beiden Fällen schneidet Dispom deutlich besser ab als die existierenden Programme.

Nach diesem eindeutigen Vergleich nutzten wir Dispom für die Erkennung von auxinabhängigen Transkriptionsfaktorbindungsstellen. Obwohl Auxin eines der wichtigsten Pflanzenhormone ist, versteht man die Bindung von auxinabhängigen Transkriptionsfaktor an die DNA nur teilweise und die Bindungsstellen werden durch eine wenig spezifischen Konsensussequenz beschrieben. Mit Dispom waren wir in der Lage ein spezifisches DNA-Bindungsmotiv zu finden, welches zudem eine hohe Positionsabhängigkeit aufweist.

Diese Ergebnisse verdeutlichen, dass die Kombination eines geeigneten Lernverfahrens und eines passenden probabilistischen Modells entscheidend für die Güte eines Programms ist.

4 Zusammenfassung

Probabilistische Modelle werden in vielen Bereichen für Datenmodellierung und Klassifikation eingesetzt. Neben der Wahl des Modells ist vor allem die Wahl des Lernverfahren kritisch, da es die Modellparameter bestimmt und damit die Güte des Modell direkt beeinflusst. In der vorgelegten Dissertation konnte ein neues Lernverfahren entwickelt werden, das einen theoretischen Rahmen für sechs bisher eher lose verbundene Lernverfahren geschaffen und eine neue Sichtweise eröffnet hat. Durch die Nutzung dieser generativen, diskriminativen und hybriden Lernverfahren konnte die Erkennung von DNA-Bindungsstellen in einer Reihe von Anwendungen signifikant verbessert.

Die im Zusammenhang mit der Dissertation erstellte quelloffene und objektorientierte Softwarebibliothek ermöglicht das einfache Kombinieren verschiedener Modelle und Lernverfahren. Zudem erlaubt es die rasche Umsetzung neuer Modelle und Lernverfahren und bildet damit die Grundlage für viele weitere Anwendungs- und Forschungsprojekte auch außerhalb der Bioinformatik.

Literatur

[AVS09] Thomas Abeel, Yves Van de Peer und Yvan Saeys. Toward a gold standard for promoter prediction evaluation. *Bioinformatics*, 25(12):i313–i320, Jun 2009.

[BEFK03] Y. Barash, G. Elidan, N. Friedman und T. Kaplan. Modeling Dependencies in Protein-DNA Binding Sites. *In proceedings of Seventh Annual International Conference on Computational Molecular Biology*, Seiten 28–37, 2003.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 1st. Auflage, 2006.

[BK97] C. Burge und S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, Apr 1997.

[BT04] G. Bouchard und Bill Triggs. The Tradeoff Between Generative and Discriminative Classifiers. In *IASC International Symposium on Computational Statistics (COMPSTAT)*, Seiten 721–728, Prague, August 2004.

[BWKT09] Jan Baumbach, Tobias Wittkop, Christiane Katja Kleindt und Andreas Tauch. Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nat Protoc*, 4(6):992–1005, 2009.

[EYSJ02] K. Ellrott, C. Yang, F. M. Sladek und T. Jiang. Identifying transcription factor binding sites through Markov chain optimization. *In Proceedings of the European Conference on Computational Biology (ECCB 2002)*, Seiten 100–109, 2002.

[Fis22] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[GKM+02] Peter Grünwald, Petri Kontkanen, Petri Myllymäki, Teemu Roos, Henry Tirri und Hannes Wettig. Supervised posterior distributions. Presented at the Seventh Valencia International Meeting on Bayesian Statistics, 2002.

[KAS+07] Tae Hoon Kim, Ziedulla K Abdullaev, Andrew D Smith, Keith A Ching, Dmitri I Loukinov, Roland D Green, Michael Q Zhang, Victor V Lobanenkov und Bing Ren. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231–1245, Mar 2007.

[KGG+08] Jens Keilwagen, Jan Grau, André Gohr, Stefan Posch und Ivo Grosse. A Java framework for statistical analysis and classification of biological sequences. http://www.jstacs.de/, 2008.

[KGR+03] Alexander E. Kel, Ellen Gössling, Ingmar Reuter, Evgeny Cheremushkin, Olga V. Kel-Margoulis und Edgar Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–3579, Jul 2003.

[LhSJR+03] Benjamin P. Lewis, I hung Shih, Matthew W. Jones-Rhoades, David P. Bartel und Christopher B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787 – 798, 2003.

[MRS+09] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas und A. G. Hatzigeorgiou. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucl. Acids Res.*, 37(suppl_2):W273–276, 2009.

[PTF+07] Heather E Peckham, Robert E Thurman, Yutao Fu, John A Stamatoyannopoulos, William Stafford Noble, Kevin Struhl und Zhiping Weng. Nucleosome positioning signals in genomic DNA. *Genome Res*, 17(8):1170–1177, Aug 2007.

[Sal97] Steven L. Salzberg. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, 13(4):365–376, 1997.

[SFMC+06] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, Annchristine Thåström, Yair Field, Irene K. Moore, Ji-Ping Z. Wang und Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, July 2006.

[SZR06] Sören Sonnenburg, Alexander Zien und Gunnar Rätsch. ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–e480, Jul 2006.

[WDKK96] E Wingender, P Dietze, H Karas und R Knuppel. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24:238–241, 1996.

[WGR+02] Hannes Wettig, Peter Grünwald, Teemu Roos, Petri Myllymäki und Henry Tirri. On Supervised Learning of Bayesian Network Parameters. Bericht HIIT Technical Report 2002-1, Helsinki Institute for Information Technology HIIT, 2002.

[YB04] Gene Yeo und Christopher B. Burge. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology*, 11(2-3):377–394, 2004. PMID: 15285897.

Jens Keilwagen wurde am 02. Oktober 1981 in Oschatz geboren. Nach dem Abitur im Jahr 2000 und anschließendem Zivildienst, studierte er von 2001 bis 2005 an Martin-Luther-Universität (MLU) Halle–Wittenberg Bioinformatik. Während seines Studiums arbeitete er am Institut für Informatik an den Lehrstühlen *Software-Engineering und Programmiersprachen*, *Theoretische Informatik* und *Datenbanken und Informationssysteme* sowie am Institut für Zoologie am Lehrstuhl *Entwicklungsbiologie*. Nach Abschluss seines Studiums begann er unter Betreuung von Prof. Grosse am IPK Gatersleben seine Promotion, die er im Juli 2010 an der MLU verteidigte. Derzeit arbeitet Jens Keilwagen am IPK Gatersleben in der Nachwuchsforschergruppe *Dateninspektion* an der Analyse von Sequenzdaten DNA-bindender Proteine und RNAs, historischen Genbankdaten und Next-Generation-Sequencing Daten.